National College of Ireland

Retail Price Optimization Using Machine Learning Algorithms

MSc Data Analytics

Research Project

Sneha Muralidhar

Student ID: x22161171

School of Computing

National College of Ireland

Supervisor: Shubham Subhnil

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sneha Muralidhar
Student ID:	X22161171
Programme:	MSc Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Shubham Subhnil
Submission Due Date:	14th December
Project Title:	Retail Price Optimization Using Machine Learning Algorithms
Word Count:	6500+
Page Count:	21

I hereby certify that the information contained in this Retail Price Optimization using Machine Learning Algorithm paper is informationpertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at therear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sneha Muralidhar
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Retail Price Optimization Using Machine Learning Algorithms

Abstract

Within the framework of a nation's economic advancement, the retail industry plays a crucial role in promoting expansion and general enhancement. Increased retailer competition highlights the significance of successful customer acquisition and retention strategies as the business landscape keeps growing. Appropriate product pricing is the cornerstone of these tactics, and it has a big impact on customer loyalty and revenue generation. Inadequate pricing optimization may result in negative financial consequences and decreased customer satisfaction. Understanding how important pricing is to retail businesses' success, this study explores price optimization as a means of striking the right balance between profitability and customer satisfaction. Using machine learning methods—in particular, decision tree regression—becomes a powerful instrument for retailers. Retailers can determine accurate pricing for their products, anticipate demand patterns, evaluate price elasticity, and create competitive pricing strategies by utilizing machine learning. Furthermore, machine learning helps with effective inventory control, guaranteeing that stores can fulfill customer demand while reducing surplus inventory. The creation of a machine learning model especially for retail price optimization is suggested by this study. By comparing pricing strategies with competitor data, forecasting demand fluctuations, and recognizing price sensitivity, the model seeks to equip retailers with data-driven decision-making capabilities to tackle pricing challenges. The model aims to provide a comprehensive solution for retailers navigating the intricacies of a constantly changing market by implementing a comprehensive approach that incorporates a variety of factors influencing retail pricing. The principal aim of the suggested machine learning model is to furnish retailers with practical insights, enabling them to make well-informed decisions grounded in a sophisticated comprehension of market dynamics. The model is a valuable asset for retailers looking to gain a competitive edge in a dynamic marketplace because of its ability to adjust to changes in competitor strategies, respond to shifts in demand, and optimize pricing for sustained profitability.

Keywords—Retail, Price Optimization, Machine Learning in retail, Competitor Analysis, Decision Tree Regressor.

1 Introduction

Retail involves selling of products to the consumers through several ways. It is one of the main sources of revenue generation in a county. Retail stands out as a cornerstone in the complex web of economic activities that characterize the prosperity of a country. Its power goes well beyond the straightforward act of purchasing and selling. Retail can take many different shapes, from the busy aisles of physical stores to the vast spaces of e-commerce, and it always creates a vibrant and diverse market. It has an impact on multiple factors such as economic, social and employment landscape of a country. Every country has small and big retail businesses. There are several retails such as E-commerce, physical stores, and more. It is a significant source of employment as it creates

opportunities for both professional and unskilled workers.

The consumers have access to wide range of goods and services through retail in an easier way. Multiple retailers sell similar products and cost of the product is one of the main factors that attracts a customer to that retail. With demand and growing technology, the competition is very high. Pricing is the important factor that attracts a customer to a business. Among the vast number of retail businesses, big and small, one basic fact remains true: price is a crucial element that can decide a company's success or failure. There are many options available to consumers, and retailer competition is intense in this age of rising demand and developing technology. Pricing must be carefully balanced; if it is too expensive, prospective sales will decline; if it is too cheap, the company may experience financial difficulties. When a product is priced too high, the business loses sales and when priced too low, they will undergo a loss.

Machine Learning models helps with predictive analysis. It can provide useful insights and helps in making right decisions using the analysis of datasets. It predicts the valuable outcome by analysing the data. In the case of retail price optimization, machine learning can help with the authentic pricing of products that in return helps with the success of a business. It helps with forecasting the demand, increasing the profitability, competitive advantage, authentic pricing, Inventory control and satisfaction of customers. A model will be built such that it takes data of several products from several different retailers and provides an outcome that will help with the price optimization suiting the profits.

This thesis' main goal is to use decision tree regression as a tool for retail price optimization, with a particular emphasis on contrasting the prices of rival retailers with the retailer's own pricing approach. The study's objective is to use this algorithm's skills to find patterns, correlations, and the best prices to meet market conditions and increase profitability. It is impossible to overestimate the importance of leveraging machine learning algorithms to optimize retail prices. Pricing has a direct impact on consumers' purchasing decisions in a time when they are more knowledgeable and discriminating than ever. A price plan that is well-optimized guarantees the shop sustainable profit margins in addition to drawing in customers. By illustrating the real-world implementation of retail price optimization, this study adds to the expanding corpus of research on the subject.

The subsequent investigation delves into the complex function of retail within a nation's economy, exploring its various manifestations and emphasizing the ubiquitous influence of pricing. The story then shifts to focus on machine learning's critical role in this environment and how it provides a potent tool for optimizing, predicting, and adjusting pricing tactics. A full grasp of how data-driven decisions and authentic pricing converge to determine success in the new retail scene emerges as we traverse the complex dance between retail and machine learning.

2 Related Work

2.1 Subsection 1

In recent years, a number of studies on price-based revenue management have been carried out. "Analytics for an Online Retailer: Price and Demand Optimization" by David Simchi-Levi, Bin Hong Alex

Lee, and Kris Johnson Ferreira (2015) [1]. The authors of this article have developed a model that primarily focuses on dynamic pricing and aids in overcoming pricing obstacles. A machine learning model that forecasts future product demand has been developed. In order to provide a pricing decision support tool for daily use, the model effectively solves the subsequent multiproduct price optimization, which combines reference price impacts. They introduced an algorithm for dynamic pricing, which modifies prices based on variables like competition and demand. In contrast, price elasticity is the primary focus of Pavithra Harsha, Shivaram Subramanian, and Markus Ettl's second work, "A Practical Price Optimization Approach for Omnichannel Retailing" (2019). They have put in place a framework for integrated data processing and machine learning that makes it possible to estimate competitive effects and location-specific, cross-channel price elasticities. In terms of determining alternative optimization models and how they apply in various scenarios, this research work is relevant to the research carried out [2].

Regression tree algorithm (2017) is used in the other work, "Demand prediction and price optimization for semi-luxury supermarket segment," to build a machine learning model that predicts weekly demand. It takes into account things like sales, vacations, and stock levels. This technique is used for pricey offline retail products. Heuristic techniques are used to improve the solution. They emphasize customized product pricing [3]. Conversely, a PEIL approach for price elasticity prediction is presented in the paper "Machine learning and operation research-based method for promotion optimization of products with no price elasticity history." It combines the nonlinear GBM model with the demand model, which employs linear relationships between the price elasticity impact and the price change ratio. This work will support the investigation with exposure to different pricing elasticity models. [4]

Steel plants' profit goals were significantly impacted when they were unable to keep up with the frequent fluctuations in the price of raw materials. By suggesting an extreme learning machine (ELM) to forecast the key raw material price in steel plants, this study seeks to allay such worries. Usually, the goal of this paper is to forecast the cost of iron ore and coking coal, which are primarily used in steel plants, by integrating Grey Relation Analysis (GRA) with a hybrid forecasting model. Here, they make an effort to set up a dynamic cost system in order to predict the cost of manufacturing finished goods and modify the production and purchasing plans accordingly. [5]

In order to predict the demand for consumer goods, this study used a traditional method called the autoregressive integrated moving average (ARIMA) model in conjunction with two machine learning techniques: artificial neural network (ANN) and support vector machine (SVM). Compared to conventional methods, such as the well-known Autoregressive Integrated Moving Average (ARIMA) model, demand forecasts using machine learning have produced better-quality results with smaller error deviations, particularly for consumer goods and complex product hierarchies. [6]

The real sales data from the retail industry was assessed by the suggested system. To assist in making purchasing decisions, the Gray Extreme Learning Machine (GELM) combines the Taguchi method with extreme learning machine and Gray Relation Analysis. With the help of GRA, one can extract the more important variables from raw data and use them as the input for a new kind of neural network, like ELM. The experimental results show that our suggested system performs better than a number of back-propagation neural network-based sales forecasting techniques, including BPN and MFLN models. [7]

Presenting a decision-support system for retail pricing and revenue optimization of these retail products is the main goal of this research (2017). This study was developed using sales data from well-known retail stores located in 45 different regions over the previous 2.5 years. Weekly demand is predicted using a machine learning algorithm based on regression trees and random forests. When making decisions, it takes into account price, holidays, sales, inventory, and other local factors. Subsequently, the interdependencies between prices and demand are measured and incorporated into an integer linear programming model to achieve the best possible price distribution. [8]

Reviewing another paper (2021) helps to study a brick-and-mortar retail price optimization problem using minimum amount of price changes and maximum number of price changes. These attributes help in efficient price optimization. Euclidean projection onto feasible region of optimization has proved efficient hence gradient projection algorithm is used. Future study of this research paper concentrates on the mathematical characteristics of the feasible set established by these two attributes as well as advanced methods that utilize those characteristics. [9]

The objective of another paper is to minimize the operative cost planning of pressurized water supply networks under credible demand forecasting over a finite time horizon. Given the difficulty of this task, it is important to use advanced mathematical procedures, which necessitates the use of well-constructed models with appropriate attributes. Based on smooth component models for the network elements, the research constructs a nonlinear mixed integer model and a nonlinear programming model with advantageous attributes for gradient-based optimization techniques. [10]

The other has demonstrated that there is a significant difference in the discussion of the static retail price optimization problem and its dynamic counterpart. While an application-focused, multidisciplinary conversation is allowed for the latter inside the framework of RM, there isn't one at the moment for the former. Rather, the subject is mostly discussed in the Marketing and Retailing literature, where a variety of demand modeling methodologies, most of which have been thoroughly investigated, have long been available to create the theoretical foundation of the static retail price optimization problem. [11]

A pricing optimization system's ability to accurately depict the price-demand connection is essential to its effectiveness. Initially, this study examines the validity of two fundamental assumptions in hotel revenue management using booking data from 28 hotels in the United States. The confluence of reduced utilization rates and increased product diversification implies that hotels had to employ distinct strategies instead of merely matching rival rates to prevent market share loss. [12]

Another paper shows that good product sales depend on a product's price being optimized. Retail pricing optimization can be enhanced by utilizing historical sales data to learn from it and come to a better conclusion. Machine learning can assist in training models on data so they can be effectively used to ascertain the optimal price for a product. A machine learning model such as random forest and multiple regression model trained with all that data may spot trends and make predictions that can be used to precisely price a model by accounting for the several elements that affect decisions. [13]

This study focuses on inventory management in retail environments, including how to handle lost sales, stock-out scenarios, and Poisson arrival rates. The retailer tries to optimize earnings and substitute costs under specific service-level limitations. The study suggests effective approximations as a solution to the problem of determining performance measurements such as predicted sales and inventory levels for many goods, which is difficult to do. A genetic algorithm is then employed to optimize order-up-to levels using these approximations. The findings show that product replacement has a beneficial effect on projected earnings and that retailers might gain from using demand substitution tactics. [14]

This study looks at a single-period inventory problem with several items, erratic demand, setup costs for production, and one-way product substitution. The problem is described as a two-stage integer stochastic program, where items are allocated to satisfy realized demand in the second stage after production quantities and products to create are decided. To create useful heuristics, the study makes use of a variety of optimization techniques, such as network flow, dynamic programming, and simulation-based approaches. The heuristics are tested computationally and compared with optimal solutions derived from large-scale linear programming. The study also provides insights into how cost parameters and demand unpredictability affect the best production and inventory decisions, emphasizing the benefits of allowing substitution. [15]

With a primary focus on manufacturers and retailers, this study examines the significance of demand data for capacity, production, and inventory planning. It draws attention to the widespread practice, despite its inherent inaccuracies of forecasting future demand using sales data. The study highlights the factors that might either lessen or increase these errors and looks at how they affect inventory costs. The replacement structure between two items during stockouts is also examined in this paper. The study uses a newsboy framework in a scenario with two products, each of which is interchangeable. To bolster its conclusions and provide guidance on ideal stocking policies, it creates sufficient and essential optimality conditions and carries out an extensive computer analysis. [16]

The Chinese real estate market, in particular the retail property values in ten major cities between 2005 and 2021, is the main topic of this study. To forecast price indices, it uses Gaussian process regression models and monthly data. Using a variety of kernels, basis functions, and predictor standardization procedures, the study investigates several model configurations and optimizes them using cross-validation and Bayesian methodologies. Using relative root mean square errors ranging from 0.0113% to 0.4835%, the study successfully creates ten accurate forecasting models.[17]

In order to forecast rent prices in Japanese apartments while taking geographical dependence into account, this study compares regression and machine learning methods. For large datasets, the regression model employs nearest neighbor Gaussian processes (NNGP), and the machine learning models include random forest (RF), deep neural networks (DNN), and XGBoost. As sample size increases, the results demonstrate that XGBoost and RF beat NNGP, with XGBoost consistently offering the highest prediction accuracy. Furthermore, the study indicates that geographical dependence can be adequately addressed in RF models by simply including spatial coordinates with the explanatory variables. [18]

The predicting of office property price indices in 10 major Chinese cities between July 2005 and April 2021 is the main objective of this study. It makes precise and understandable forecasts by using neural networks (NNs). The study investigates several model configurations that take data-splitting ratios, delays, algorithms, and hidden neurons into account. The results show that a simple neural network (NN) with three delays and three hidden neurons may achieve stable performance during training, validation, and testing, with an average relative root mean square error of roughly 1.45% across the ten cities. These findings can be utilized separately or in conjunction with fundamental projections to shed light on trends in office property prices and guide policy decisions. [19]

The weekly wholesale price index of edible oil in the Chinese market from January 1, 2010, to January 3, 2020, is the main topic of this study, which addresses the need for price projections in the agricultural market. The study uses a non-linear auto-regressive neural network as its forecasting model and investigates different model configurations, such as data segmentation, hidden neurons, training techniques, and delays. Based on the findings, a reasonably simple model with accurate and consistent performance is developed. 2.80%, 3.01%, and 1.80%, respectively, are the relative root mean square errors for training, validation, and testing. In policy analysis pertaining to agricultural commodities, these forecast results may supplement other fundamental forecasts and be useful for technical analysis. [20]

Investment models are being transformed by the incorporation of artificial intelligence (AI) tools into urban real estate decision support systems, which is being driven by increasing digitization. Examining seventy pertinent papers from the Scopus database, this study finds a concentration of publications in nations with high levels of digitalization. Despite this, research frequently uses limited data sets and favors straightforward machine learning techniques over deep learning. Subsequent initiatives ought to capitalize on the increasing digitization of urban areas, stressing the necessity of more extensive and precise statistics. Resolving model explainability issues is essential for real estate stakeholders to use decision support systems effectively. [21]

3 Methodology

This section describes the methodical approach used to accomplish the study's goals, with a particular emphasis on the use of decision tree regression for retail pricing optimization based on competition data.

A. Data

i. Data Collection

Choosing the right dataset is the foundation of a study. A comprehensive dataset that includes historical retail pricing, competition prices, production costs, promotional activities, and other pertinent variables forms the basis of the analysis. To guarantee representativeness and diversity, data were obtained from several retail locations within the selected industry. Analysis of seasonality and temporal trends is possible because the dataset covered a predetermined time frame.

The first stage of this study was to collect a rich and varied dataset from the Kaggle platform, which is

a well-known open-source repository for datasets from several fields. With its vast array of datasets offered by scholars, professionals, and hobbyists, Kaggle served as a useful tool for gathering information about the dynamics of retail pricing. A crucial choice was made while selecting the Kaggle dataset, and it was based on how pertinent the information was to the study's goals. To make sure the study could adequately reflect the intricacies of retail pricing optimization, a dataset comprising competitor prices, historical retail prices, and related variables was used.

ii. Dataset Characteristics

The selected dataset contained a wide range of features, such as but not restricted to:

- Competitor costs for similar goods.
- Distribution of total price by holidays
- Distribution of total price by weekend
- Distribution of total price by weekday
- Average total price by product category

The depth and breadth of the dataset were designed to offer a thorough understanding of the variables affecting retail prices, facilitating an in-depth investigation of the decision tree regression model's potential.

iii. Data Cleaning and Preprocessing:

Kaggle datasets frequently present unique difficulties, such as inconsistent data, outliers, or missing values. To guarantee the quality and dependability of the dataset, thorough preprocessing and data cleaning were carried out. To avoid skewed influences on the model, missing values were imputed or handled appropriately, outliers were dealt with, and data normalization was carried out.

iv. Ethical and Legal Compliance:

Throughout the data collection process, it was crucial to follow the law and ethical guidelines. User agreements and licensing requirements are frequently included with Kaggle datasets; these were thoroughly examined and complied with. One of the most important aspects of the ethical considerations in the research was ensuring data privacy and anonymity, especially when working with potentially sensitive information.

The Kaggle platform was the focal point of the data collection phase, which provided the foundation for the rest of the research. By utilizing the various carefully selected datasets on Kaggle, the research sought to derive significant understandings into the complex dynamics of retail pricing. Thorough preprocessing, augmentation, and data cleaning procedures were used to improve the quality and richness of the dataset, guaranteeing that the decision tree regression model would have a strong basis on which to analyse and optimize retail prices. The data collection process was designed with ethical considerations in mind, guaranteeing a responsible and transparent use of the open-source data that was acquired from Kaggle.

B. Exploratory Data Analysis (EDA)

EDA was used to learn more about the distribution of the variables, spot possible trends, and comprehend the dynamics of rival pricing in connection to the retailer's pricing plan. Scatter plots and correlation matrices were two examples of visualization techniques used to identify important relationships between variables and provide guidance for further modeling decisions.

C. Model Selection

Decision Tree Regression can handle nonlinear relationships and produce results that are comprehensible, hence it is selected as the main modeling technique. This algorithm is chosen because it was deemed appropriate for capturing the intricate relationships present in retail pricing dynamics. The decision tree regression model was chosen after a comparative analysis with other models, including support vector machines and linear regression.

i. Non-linear Connections:

Nonlinear relationships between a variety of factors, including rival prices, production costs, and promotional activities, frequently have an impact on retail pricing. Decision trees are an excellent choice for modeling the complex and dynamic nature of retail pricing because they possess an innate ability to detect nonlinear patterns in the data.

ii. Interpretability:

An understandable and transparent depiction of the decision-making process is offered by decision trees. This is especially important in the retail industry, where stakeholders must comprehend both the forecasts and the contributing factors. Decision trees' interpretability makes it easier to identify important factors and how they affect pricing decisions, providing insights that can be put to use.

iii. Handling Complex Interactions:

Retail pricing involves intricate relationships between a number of variables, which decision trees are particularly good at managing. Decision trees' hierarchical structure enables them to identify and rank the importance of various features, even those that interact in complex ways. This is necessary to comprehend how, for example, adjustments to competitor pricing may interact with other factors to affect ideal retail prices.

iv. Adaptability to Change:

Due to the dynamic nature of retail markets, pricing strategies frequently need to be adjusted. Decision trees are robust in situations where market dynamics change because they can adjust to changes in the data. For retail price optimization models to react to changes in consumer demand, rival behavior, and outside economic factors, this flexibility is essential.

v. Handling Heterogeneous Data:

Retail datasets may contain both numerical and categorical features, among other types of variables. Decision trees don't require a lot of pre-processing in order to handle mixed data types and a variety of input feature sets. This adaptability is useful for handling the wide variety of data found in retail datasets.

vi. Sensitivity testing and scenario analysis:

Retailers can simulate various pricing scenarios by using decision tree models, which are especially well-suited for scenario analysis and sensitivity testing. This is especially helpful for retail price optimization, where stakeholders must assess the potential effects of changes in competitor pricing or other variables on recommendations for the best prices. Decision trees offer a structure for carrying out these kinds of evaluations.

vii. Collaborative Techniques for Enhanced Achievement:

Regressors for decision trees can be improved by using ensemble techniques like Gradient Boosting and Random Forests. Several decision trees are combined in these methods in order to increase generalizability and prediction accuracy. Ensemble approaches can further improve the model's performance in the context of retail price optimization, resulting in more reliable and accurate pricing recommendations.

While decision tree regressors have many benefits for retail price optimization, it's important to remember that the best algorithm will depend on the particulars of the dataset and the intricacies of the particular problem. To determine the best course of action for a specific retail pricing scenario, it is frequently beneficial to compare the performance of decision tree regressors with that of other regression models.

D. Model Training

To make training and evaluating the model easier, the dataset was divided into training and validation sets. Using the training set, the decision tree regression model was trained, and its performance was optimized through hyperparameter tuning. In order to balance model complexity and generalizability, parameters like tree depth and minimum sample split had to be adjusted iteratively.

E. Model Evaluation

To evaluate the model's performance in an out-of-sample situation, the validation set was employed. Metrics such as Mean Squared Error (MSE) and R-squared were utilized to assess the predictive precision of the model.

4 Design Specification

The design specification lists the essential components and factors to be taken into account when putting into practice a decision tree regressor model made especially for retail price optimization. This covers feature selection, model hyperparameter adjustments, data preprocessing techniques, and approaches to addressing difficulties particular to the retail industry and provides a clear understanding of the design.

A. Feature Selection:

Total Price:

This is the target variable that the model is attempting to optimize. The product's prices based on the quantity serve as the foundation for understanding past pricing trends and predicting future optimal prices.

Competitor price:

For the decision tree regressor to learn and modify pricing strategies based on the competitive landscape, competitor prices must be incorporated. It facilitates the model's comprehension of market dynamics and allows it to modify the retailer's pricing strategy as necessary.

Lag price:

This is one of the main attributes that helps the model understand price difference between the products. Model can capture the lag price and adjust the prices accordingly.

B. Data Preprocessing:

Handling missing values:

If there are any missing values in the dataset, they must be fixed. To maintain the dataset's integrity, imputation techniques or the removal of rows or columns with missing values should be taken into account. The dataset used for the model development does not contain any missing values.

Normalization:

Normalization is not important in this model as the features in the dataset are naturally on a similar scale and there are no features with noticeably different scales. Normalization might not have as much of an effect as all features are already expressed in the same unit.

C. Model Hyperparameters:

Tree Depth:

The decision tree's depth controls the model's complexity. To avoid either an underfit or an overfit, it should be carefully adjusted. The ideal tree depth can be found by using cross-validation techniques.

Minimum Sample Split:

It is ensured that nodes in the decision tree are split only when a specific number of samples are present by setting an appropriate minimum sample split. This enhances generalizability and controls the tree's granularity.

Minimum Leaf Samples:

The granularity of the terminal nodes is influenced by the minimum number of samples needed to form a leaf node. This hyperparameter can be changed to affect how sensitive the model is to noise.

D. Evaluation Metrics:

A set of evaluation metrics, including the R-squared (R2) score and Mean Absolute Error (MAE) were used to rigorously assess the performance of the decision tree regressor.

R-squared (R2) Score:

The percentage of the target variable's variance that can be predicted from the independent variables is indicated by the R2 score. A high R2 score means that the retail price variability is well captured by the model.

Mean Absolute Error (MAE):

The average absolute difference between expected and actual prices is quantified by MAE. A lower MAE implies that the model predicts prices accurately and with few errors.

A detailed implementation blueprint for a decision tree regressor specifically suited for retail price optimization is provided by this design specification. The model that is produced is ready to provide practical insights and improve pricing strategies in the competitive and ever-changing retail industry because it takes into account certain features, preprocessing procedures, model hyperparameters, and ethical considerations.

5 Implementation

The retail price optimization model's implementation phase entails translating the design specification into executable code. This section describes the steps involved in coding and running the decision tree regressor, including data preprocessing, model evaluation and building predictive models for the 'total_price' variable.

i. Importing Libraries

Importing the Python libraries required for machine learning, data manipulation, and visualization is the first step in the implementation.

ii. Loading Dataset

Loading the retail pricing dataset obtained from Kaggle.

iii. Data Preprocessing

Data is then checked for null values. For the used dataset, there are no null values found.

```
print(dataset.isnull().sum())
```

```
product_id                      0
product_category_name           0
month_year                      0
qty                             0
total_price                     0
freight_price                   0
unit_price                      0
product_name_lenght             0
product_description_lenght      0
product_photos_qty              0
product_weight_g                0
product_score                   0
customers                       0
weekday                         0
weekend                         0
holiday                         0
month                           0
year                            0
s                               0
volume                          0
comp_1                          0
ps1                             0
fp1                             0
comp_2                          0
ps2                             0
fp2                             0
comp_3                          0
ps3                             0
fp3                             0
lag_price                       0
dtype: int64
```

Fig 1: Dataset checked for null values.

iv. Explanatory Data Analysis

EDA was used to learn more about the distribution of the variables, spot possible trends, and comprehend the dynamics of rival pricing in connection to the retailer's pricing plan. Below are some of the graphs plotted to understand the pricing strategy of the retailer.

The below graph shows the total price distribution for the products.

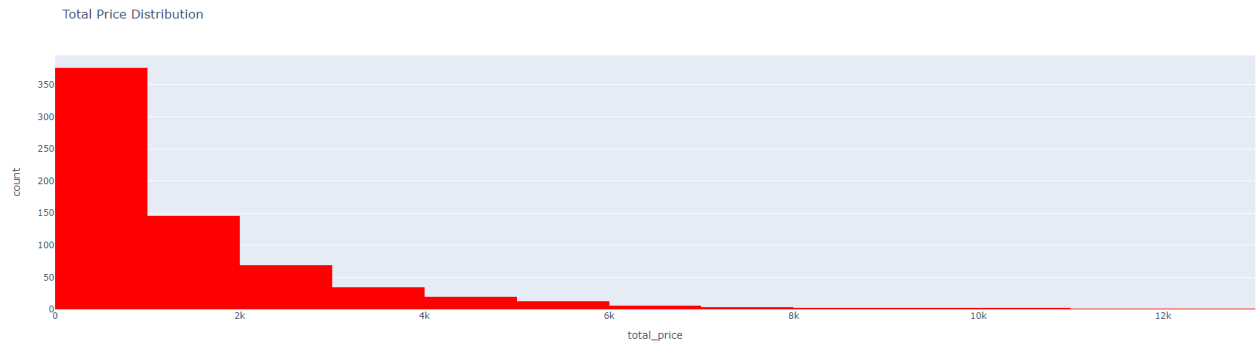

Fig 2: Total distribution of products.

The below graph shows the unit price distribution for the products.

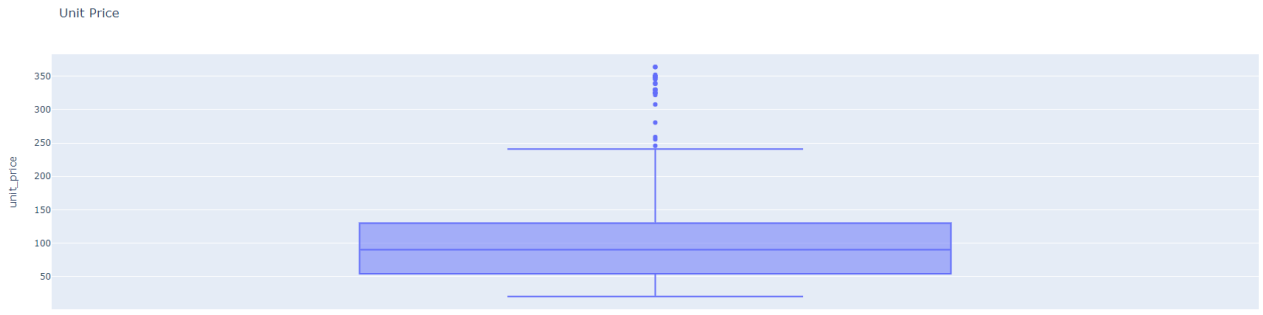

Fig 3: Unit price distribution.

The below scatter plot shows total price vs quantity of products.

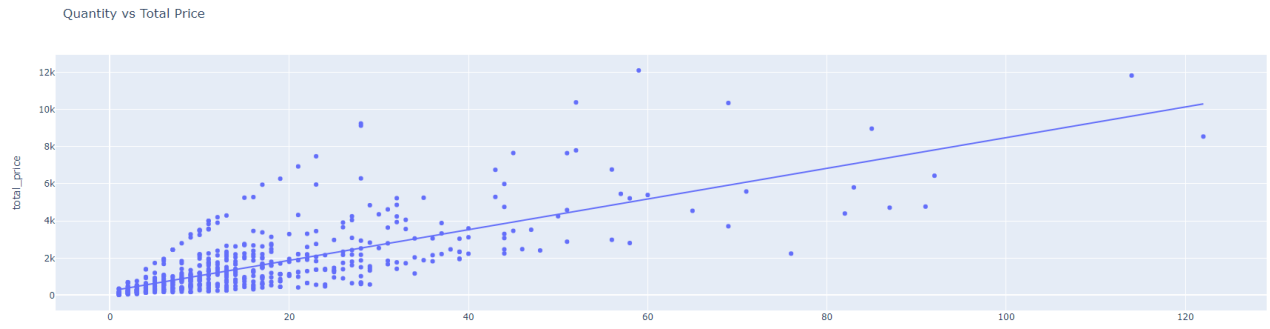

Fig 4: Quantity vs Total Price.

The bar graph shows Average Total Price by Product Category.

Fig 5: Average Total Price by Product Category.

A correlation matrix is shown below for the numerical features.

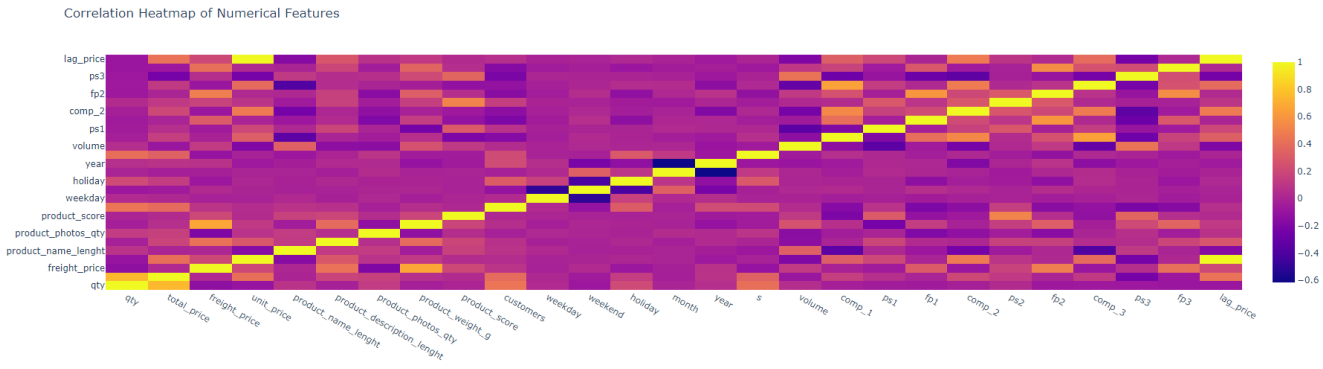

Fig 6: Correlation HeatMap.

Below graph shows Average Competitor Price Difference by Product Category.

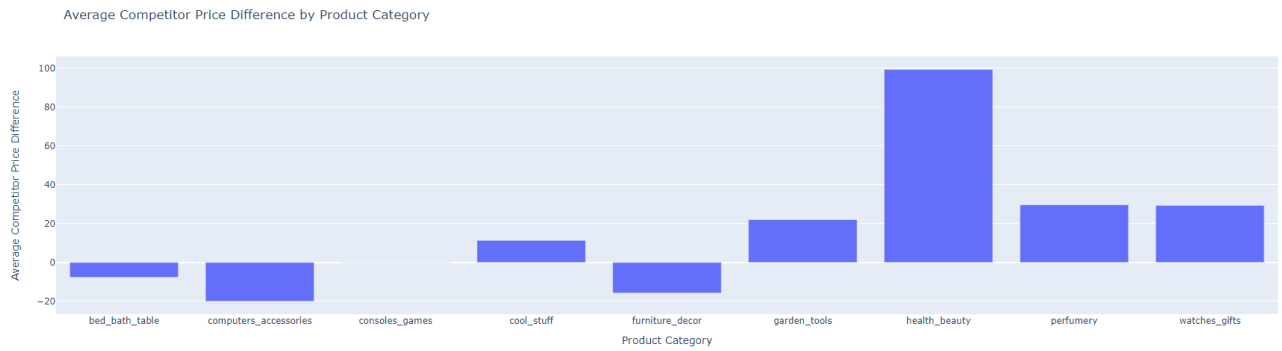

Fig 7: Average Competitor Price Difference by Product Category.

v. Feature Selection

The feature matrix 'x' contains independent variables such as 'qty', 'unit_price', 'comp1', 'product score', 'comp_price_diff' which are used for training the machine learning model. The target variable 'y' is the variable to be predicted. Here 'total_price' is the target variable.

```
X = dataset[['qty', 'unit_price', 'comp_1',
          'product_score', 'comp_price_diff']]
y = dataset['total_price']
```

Fig 8: Feature Selection.

vi. Data Splitting

Dataset is then split into train and test data to train the model to achieve maximum accuracy.

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.2,
                                                    random_state=42)
```

Fig 9: Data Splitting.

vii. Model Training

The Decision Tree Regressor model is then trained on Training set.

```
model = DecisionTreeRegressor()
model.fit(X_train, y_train)
```

Fig 10: Model Training.

viii. Model Prediction

Predict the values based on validation set.

```
y_pred = model.predict(X_test)
```

Fig 11: Model Prediction.

ix. Evaluation Metrics

Evaluating model's performance using the specified evaluation metrics.

```
r2 = r2_score(y_test, y_pred)
print(f"R-squared (R2) score: {r2}")

# Calculate the Mean Absolute Error (MAE)
mae = mean_absolute_error(y_test, y_pred)
print(f"Mean Absolute Error (MAE): {mae}")
```

```
R-squared (R2) score: 0.9532337391390635
Mean Absolute Error (MAE): 152.7873529411765
```

Fig 12: Evaluation of model.

The design specification is turned into a functional decision tree regressor model for retail price optimization during the implementation phase. Through the processes of loading the dataset, choosing features, preprocessing the data, and training the model, stakeholders are able to obtain important insights into pricing dynamics. Evaluation metrics and visuals offer a thorough picture of the model's performance. Frequent modifications and enhancements guarantee that the model stays flexible in response to evolving market circumstances and keeps supporting the best possible retail pricing tactics.

6 Evaluation

R2 Score:

The R2 score indicates how well the decision tree regressor explains the variation in actual prices. A score close to 1 indicates a good fit, indicating that the model captures a significant portion of the data's variability.

R2 Score is calculated as follows:

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sum of Squares}}$$

A high R2 score validates the model's ability to predict optimal prices based on the features taken into account.

The R2 value for model 1 is 0.95 and for model 2 is 0.94. R2 values is close to 1 which indicates that the model perfectly explains the variability in the target variable.

Mean Absolute Error (MAE):

The average absolute difference (MAE) between the actual and predicted prices is shown. It offers an indicator of how well the model predicts pricing values. A lower mean absolute error (MAE) suggests that the decision tree regressor minimizes the overall pricing error by producing accurate predictions.

MAE is calculated as shown below:

$$MAE = \frac{\sum|\text{Actual Price} - \text{Predicted Price}|}{\text{Number of Observations}}$$

The MAE value for model 1 is 152.7 which means that the model prediction is off by 152.7 units than the actual values and for model 2 it is 157.7.

Results:

Final graph is plotted which shows the actual values needed to maximize profits by considering competitor prices. Below graph shows predicted vs actual retail price.

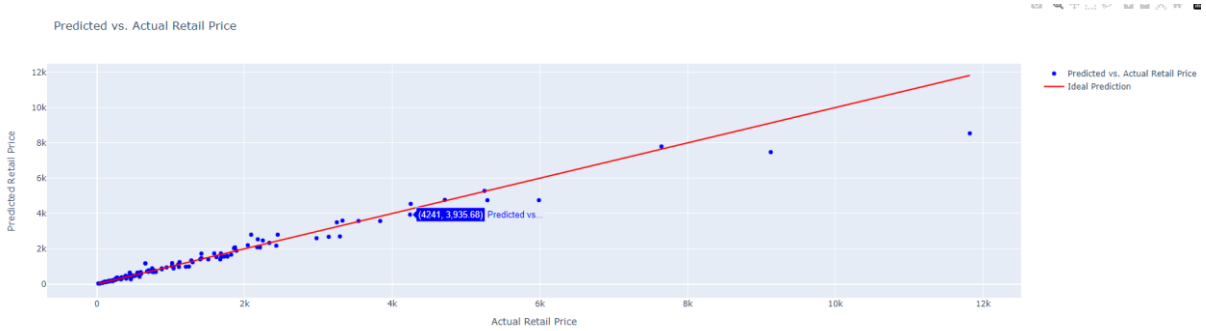

Fig 13: Predicted vs Actual Retail Price for Model 1.

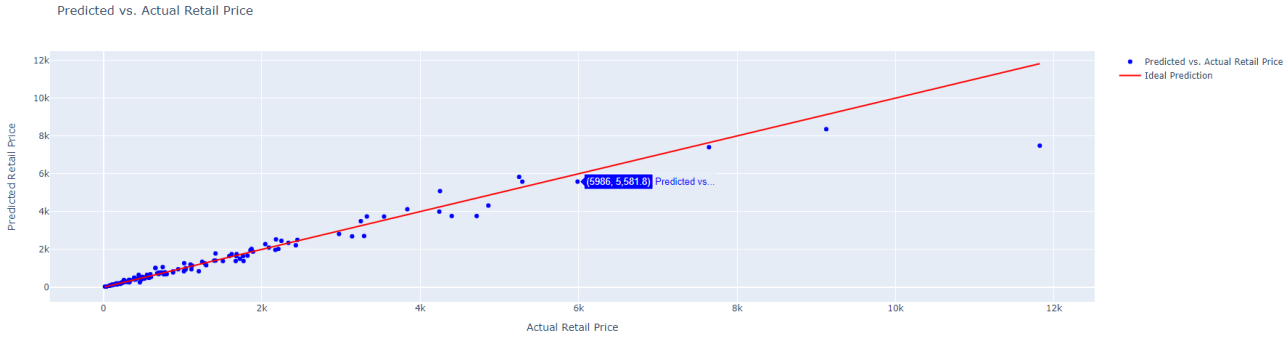

Fig 14: Predicted vs Actual Retail Price for Model 2.

6.1 Case Study 1: Default Hyperparameters (Model 1)

Introduction:

In Case Study 1, the performance of a Decision Tree Regressor was evaluated using default hyperparameters on our dataset. The model aimed to predict the 'total_price' based on several features.

Model Specifications:

Algorithm: Decision Tree Regressor

Hyperparameters: Default settings

Results:

R^2 Score: 0.95

Mean Absolute Error (MAE): 152.3

The model with default hyperparameters performs well in terms of prediction. The high R2 score of 0.95 indicates that the model explains approximately 95% of the variance in the target variable. Furthermore, the low MAE of 152.3 indicates that the model's predictions are on average 152.3 units off from the actual values.

6.2 Case Study 2: Customized Hyperparameters (Model 2)

Introduction:
In Case Study 2, the impact of adjusting hyperparameters on the performance of the Decision Tree Regressor was explored. The model used specific hyperparameter settings to optimize its predictive capabilities.

Model Specifications:

Algorithm: Decision Tree Regressor

Hyperparameters:

Criterion: Squared error

Splitter: Best split

Max Depth: 10

Min Samples Split: 5

Min Samples Leaf: 2

Max Features: All features considered.

Random State: 42

Results:

R^2 Score: 0.94

Mean Absolute Error (MAE): 157.7

Model 2 shows good predictive performance as well, with adjusted hyperparameters. With an R2 score of 0.94, the target variable's variance is roughly 94% explained by the model. Despite being marginally greater than Model 1, the MAE of 157.7 indicates precise forecasts, with an average deviation of 157.7 units from the observed values.

6.3 Discussion

To sum up, Case Studies 1 and 2 demonstrate the Decision Tree Regressor's capacity to accurately predict 'total_price' within our dataset. With default hyperparameters, Case Study 1 produces a low Mean Absolute Error (MAE) of 152.3 and an impressive R2 score of 0.95, demonstrating strong predictive performance. In Case Study 2, the model obtained a marginally higher MAE of 157.7 and a slightly lower R^2 score of 0.94 after hyperparameters were customized for optimization. The comparative analysis highlights the significance of carefully considering the trade-offs associated with hyperparameter tuning and indicates that default settings might be appropriate for our dataset. Even

though both models show good predictive power, more research could improve overall performance and improve model interpretability. This research could include feature importance analysis and possible ensemble methods. Price prediction benefits greatly from the use of the Decision Tree Regressor, and further improvement could yield even more insights for the best possible model deployment.

7 Conclusion and Future Work

An important step toward improving pricing strategies in the retail industry has been taken with the adoption of the decision tree regressor for retail price optimization to answer the research question "**What elements and strategies can be used to optimize retail pricing to increase profitability, improve customer satisfaction, and keep a competitive advantage?**". The model employs a methodical approach that utilizes past data, competitor pricing dynamics, and pertinent features to suggest prices that are in line with market conditions. The model's output, as demonstrated by assessment metrics and visualizations, offers insightful information about how effective the strategy is.

The decision tree regressor's capacity to recognize nonlinear relationships and adjust to shifting market conditions is one of its main advantages. The model's transparency makes it easier to understand and helps stakeholders make well-informed decisions about pricing strategies. The impact of the model's recommendations over time can be easily visualized with the help of the graphical representation of actual prices, competitor prices, and optimized prices.

The design incorporates ethical considerations, including privacy protection, transparency, and fairness, to emphasize the responsible use of machine learning and data in the retail industry. Upholding ethical standards is crucial because retail pricing is a delicate topic that directly impacts customers.

It is important to understand that although the current implementation offers a basis for retail price optimization, the model can be improved and expanded upon in subsequent iterations. Maintaining the model's relevance and efficacy will require ongoing performance evaluation, stakeholder input, and integration of new retail trends.

Future Work

The project lays the groundwork for upcoming initiatives to improve the retail price optimization model even more. Future research and development could focus on a number of areas, such as:

1. **Ensemble Methods:**
   Examine using ensemble techniques to enhance the predictive accuracy and robustness of the model, such as Gradient Boosting or Random Forests. The use of ensemble methods can improve generalizability and reduce overfitting.

2. **Dynamic Feature Integration:**
   Examine how other dynamic features, like metrics for consumer behavior, economic

indicators, or outside market trends, might be integrated to affect pricing. The capacity of the model to capture complex pricing dynamics can be improved by extending its consideration to a wider range of variables.

3. **Fine-Tuning Hyperparameters:**
   Adjust the decision tree regressor's parameters further by carrying out a more thorough hyperparameter tuning procedure. The model's performance can be further enhanced by optimizing parameters like minimum leaf sample size, minimum sample split, and tree depth.

4. **Incorporating External Data Sources:**
   Add to the dataset external data sources that offer supplementary details about industry trends, market conditions, or geopolitical factors. Incorporating outside data can help develop a more thorough understanding of the variables affecting retail prices.
5. **Real-time Updates:**
   Implement real-time updates and continuous learning mechanisms. Retail markets are dynamic, and the ability to adapt the model in real-time to changing conditions improves its responsiveness and effectiveness.

6. **Explainability Enhancements:**
   Increase the model's explainability by adding more interpretability tools or strategies. A better understanding of the model's pricing process can help stakeholders have more faith in one another.

7. **Scalability Considerations:**
   Examine the model's scalability to make sure it can handle growing datasets and changing business requirements. The more products or markets the model is applied to, the more important scalability considerations become.

The retail price optimization model can develop into a more advanced and flexible tool by taking into account these future work considerations. This will benefit retailers who are looking to maximize their pricing strategies in a dynamic and competitive marketplace.

References

[1] Kris Johnson Ferreira, Bin Hong Alex Lee, David Simchi-Levi (2015). "*Analytics for an Online Retailer: Demand Forecasting and Price Optimization*". Manufacturing & Service Operations Management 18(1):69-88.

[2] Pavithra Harsha, Shivaram Subramanian, Markus Ettl (2019), "*A Practical Price Optimization Approach for Omnichannel Retailing*". INFORMS Journal on Optimization 1(3):241-264.

[3] Qu, T.; Zhang, J.H.; Chan, F.T.; Srivastava, R.S.; Tiwari, M.K.; Park, W.-Y. "*Demand prediction and price optimization for semi-luxury supermarket segment*". Comput. Ind. Eng. 2017, 113, 91–102.

[4] Greenstein-Messica, A.; Rokach, L. "*Machine learning and operation research based method for promotion optimization of products with no price elasticity history*". Electron. Commer. Res. Appl. 2020, 40, 100914.

[5] Tsung-Yin Ou, Chen-Yang Cheng, Po-Jung Chen, Chayun Perng, "*Dynamic cost forecasting model based on extreme learning machine - A case study in steel plant*". Computers & Industrial Engineering. 2016, 544 – 553.

[6] Kandananond, K, "*A comparison of various forecasting methods for autocorrelated time series*", International Journal of Engineering Business Management. 2012, 1.

[7] F.L. Chen, T.Y. Ou, "*Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry*", Expert Systems with Applications. 2011, 1336-1345.

[8] Qu, T., Zhang, J. H., Chan, F. T. S., Srivastava, R. S., Tiwari, M. K., & Park, W.-Y, "*Demand prediction and price optimization for semi-luxury supermarket segment*", Computers & Industrial Engineering. 2017

[9] Wang, Xiaojie & Huang, Hsin-Chan & Han, Lanshan & Lim, Alvin. (2021). "*Price Optimization with Practical Constraints*".

[10] J. Burgschweiger, B. Gn"adig, and M. C. Steinbach. "*Optimization models for operative planning in drinking water networks*". Optimization and Engineering, 10:43-73, 2009

[11] T. P. Kunz and S. F. Crone. "*Demand models for the static retail price optimization problem - a revenue management perspective*". SCOR, pages 101-125, 2014.

[12] S. Lee. "Study of demand models and price optimization performance". PhD thesis, Georgia Institute of Technology, 2011.

[13] Subbarayudu, Y., Reddy, G.V., Raj, M.V.K., Uday, K., Fasiuddin, M.D. and Vishal, P., 2023. "An efficient novel approach to E-commerce retail price optimization through machine learning". In E3S Web of Conferences (Vol. 391, p. 01104). EDP Sciences.

[14] Tan, Baris & Karabati, Selcuk. (2013). "Retail inventory management with stock-out based

dynamic demand substitution". International Journal of Production Economics. 145. 78-87. 10.1016/j.ijpe.2012.10.002.

[15] Rao, Uday & Swaminathan, Jayashankar & Zhang, Jun. (2004). "Multi-product inventory planning with downward substitution, stochastic demand and setup costs". IIE Transactions. 36. 59-71..

[16] Ernst, Ricardo & Kamrad, Bardia. (2006). "Estimating demand by using sales information: Inaccuracies encountered". European Journal of Operational Research. 174. 675-688.

[17] Xu, Xiaojie & Zhang, Yun. (2023). "A Gaussian process regression machine learning model for forecasting retail property prices with Bayesian optimizations and cross-validation". Decision Analytics Journal.

[18] Yoshida, Takahiro & Murakami, Daisuke & Seya, Hajime. (2022). "Spatial Prediction of Apartment Rent using Regression-Based and Machine Learning-Based Approaches with a Large Dataset". The Journal of Real Estate Finance and Economics.

[19] Xu, Xiaojie & Zhang, Yun. (2023). "Office property price index forecasting using neural networks". Journal of Financial Management of Property and Construction.

[20] Xu, Xiaojie & Zhang, Yun. (2023). "Edible oil wholesale price forecasts via the neural network". Energy Nexus. 12. 100250.

[21] Koumetio Tekouabou, Cédric Stéphane & Gherghina, Ştefan Cristian & Kameni, Eric & Filali, Youssef & Idrissi Gartoumi, Khalil. (2023). "AI-Based on Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey". Archives of Computational Methods in Engineering.