# 

**Project**

**“Determining Optimal Pricing for Retail Products”**

**Submitted By,**

Mavani Krishnkumar Manjibhai

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr.no** | **Topic Name** | **Page.no** |
| **1** | **Project Introduction** | **3** |
| 1.1 | Domain | 3 |
| 1.2 | Definition | 3 |
| 1.3 | Problem Addressing | 3 |
| 1.4 | Benefits | 3 |
| 1.5 | Issues | 4 |
| 1.6 | Challenges | 4 |
| 1.7 | Objective | 5 |
| 1.8 | Summary | 5 |
| **2** | **Design Details** | **6** |
| 2.1 | Architecture | 6 |
| 2.1.1 | Data Flow Architecture | 6 |
| 2.1.2 | Exploratory Data Analysis (EDA) & Visualization | 7 |
| 2.1.3 | Feature Engineering | 7 |
| 2.1.4 | Model Flow Architecture | 7 |
| 2.1.5 | Visualization Flow Architecture | 9 |
| 2.1.6 | Model Deployment | 9 |
| 2.1.7 | Complete Flow Diagram | 10 |
| 2.1.8 | Implementation Workflow | 11 |
| 2.2 | Dataset Details | 13 |
| **3** | **Implementation Details of Project** | **15** |

1. **Project Introduction**
   1. **Domain**

The domain of this project is **Retail and E-Commerce Analytics**, with a specific focus on price optimization using machine learning techniques. This domain involves analyzing historical sales data, competitor pricing, and product attributes to determine optimal price points. By leveraging predictive modeling and data-driven insights, businesses can enhance profitability, maintain competitive positioning, and refine their pricing strategies to align with market dynamics.

This project encompasses various aspects of **data science and machine learning** to develop an intelligent pricing model. It integrates **time-series forecasting, feature engineering, and model interpretability** to provide actionable recommendations. By employing regression models and explainability techniques, the solution enables retailers to make informed pricing decisions, ensuring revenue maximization while adapting to customer demand patterns and seasonal trends.

* 1. **Definition**

The project aims to develop a machine learning-based price optimization model that analyzes historical sales data, competitor pricing, and product attributes to determine optimal pricing strategies for retail products. By leveraging predictive analytics and regression models, the solution will identify price points that maximize profitability while maintaining market competitiveness.

* 1. **Problem**

The problem is a **supervised regression** problem aimed at optimizing retail product pricing. The goal is to predict the optimal price for a product based on historical sales data, competitor pricing, and other relevant factors. This problem falls under supervised learning, as the model is trained on labeled data containing past pricing and sales performance, allowing it to learn patterns and relationships. By leveraging this trained model, the system can generate accurate price recommendations for new or existing products to maximize profitability while maintaining market competitiveness.

* 1. **Benefits**
* **Profit Maximization**: The model will identify optimal price points that maximize revenue and profit margins while maintaining competitive pricing.
* **Market Competitiveness**: By analyzing competitor pricing trends, the company can strategically position its products to attract customers without compromising profitability.
* **Data-Driven Decision Making**: Leveraging historical sales data, customer behavior, and seasonal trends, the model will provide actionable insights for informed pricing strategies.
* **Demand Forecasting**: Predicting optimal pricing based on past sales performance enables better inventory and sales planning, reducing overstock and lost sales opportunities.
* **Operational Efficiency**: Automating the pricing optimization process minimizes manual effort, ensuring consistent and scalable pricing strategies across product categories.
  1. **Issues**
     + 1. **Data Quality and Completeness**: Inconsistent, missing, or inaccurate sales and pricing data can affect the model’s performance and reliability.
       2. **Competitor Pricing Volatility**: Rapid and unpredictable changes in competitor pricing can make it difficult for the model to adapt in real time.
       3. **Consumer Behavior Variability**: Fluctuations in customer demand due to seasonal trends, promotions, or external economic factors can impact pricing effectiveness.
       4. **Model Interpretability**: Complex machine learning models, such as ensemble methods, may lack transparency, making it challenging for businesses to trust and implement pricing recommendations.
       5. **Regulatory and Ethical Considerations**: Price optimization must comply with fair pricing laws and ethical standards to avoid price discrimination or anti-competitive practices.
  2. **Challenges**
     + 1. **Feature Selection and Engineering**: Identifying the most influential features, such as competitor pricing, demand elasticity, and product attributes, requires extensive experimentation and domain expertise.
       2. **Balancing Profitability and Competitiveness**: Setting optimal prices that maximize profits without losing customers to competitors is a delicate balance.
       3. **Handling Dynamic Market Conditions**: The model must be adaptable to changing market trends, new product launches, and unexpected economic shifts.
       4. **Scalability of the Model**: Ensuring that the model can handle large datasets across multiple product categories while maintaining efficiency is critical.
       5. **Real-Time Pricing Adjustments**: Implementing real-time or near-real-time price adjustments based on model predictions requires robust infrastructure and automation capabilities.
  3. **Objective**

The objective of this project is to develop a machine learning-based price optimization model that leverages historical sales data, competitor pricing, and market trends to determine optimal product pricing. The model aims to maximize profitability while ensuring competitive positioning through data-driven insights. Additionally, it enhances decision-making by providing accurate price recommendations that adapt to changing market conditions.

* 1. **Summary**

This project focuses on developing a machine learning-driven price optimization model to enhance retail pricing strategies. By analyzing historical sales data, competitor pricing, and market trends, the model predicts optimal price points to maximize profitability while maintaining a competitive edge. Through data-driven insights, the solution enables businesses to make informed pricing decisions, improving revenue generation and operational efficiency.

1. **Design Details**
   1. **Architecture**The architecture of the **Retail Price Optimization System** follows a structured, modular approach to ensure an efficient and systematic workflow for data analysis and model development. The key stages include:
      1. **Data Flow Architecture**
2. **Import Necessary Libraries:**
   * **Objective:** Import essential libraries for data manipulation, preprocessing, visualization, model training, and evaluation.
   * **Libraries:** pandas, numpy, matplotlib, seaborn, sklearn, shap, joblib, warnings, etc.
3. **Load the Dataset:**
   * **Objective:** Load the dataset into a pandas DataFrame and create a copy for further exploration and processing.
   * **Implementation:** Use pd.read\_csv() to load the dataset ‘**retail\_price.csv**’ and store it in the old\_df variable. Copy data from old\_df to df for further processing.
4. **Data Exploration:**
   * **Objective**: Understand the data structure in the dataset.
   * **Tasks**:
     1. Display shape of the dataset (df.shape).
     2. Dataset in find all column name (df.,columns)
     3. Show the first 5 rows (df.head()).
     4. Summarize data (df.info()).
5. **Data Preprocessing (Data Preparation):**
   * **Objective**: Clean and preprocess the data to make it suitable for machine learning.
   * **Tasks**:
     1. Handle Missing Values
        + Identify columns with missing values and their count.
        + Visualize missing values using a heatmap.
     2. Handle Duplicates (Identify and count duplicate rows).
     3. Identify Outliers in Every Columns.
     4. Identify Unique columns in the data.
     5. Identify categorical features & numerical features columns.
     6. Perform Statistical Operation in numeric columns
     7. **Exploratory Data Analysis (EDA) & Visualization**
        + - Distribution of Total Price, Unit Price, Freight Price, Product Weight, Customers per Product, Product Ratings, Product Prices
          - Trend Over Time (Sales Volume, Total Price)
          - Top 10 Product Categories by Sales Volume
          - Product Category vs. Average Total Price
          - Monthly Trend of Freight Price & Product Score
          - Heatmap of correlation matrix ('product\_weight\_g', 'qty', 'total\_price', 'freight\_price', 'unit\_price', 'product\_score', 'customers', 'weekday', 'weekend', 'holiday')
          - Pairplot ('product\_weight\_g', 'qty', 'total\_price', 'freight\_price', 'unit\_price', 'product\_score', 'customers', 'weekday', 'weekend', 'holiday')
     8. **Feature Engineering**
6. Calculate Revenue, Profit, Margin
7. Price Ratio
8. Price Differences
9. Market Demand Indicators
10. Time-related Features
11. Temporal features
12. Product attribute ratios
13. Define input features (X) and target variable (y).
14. Split data into train and test sets (80% train, 20% test) using train\_test\_split().
15. Standardize the features using StandardScaler.
    * 1. **Model Flow Architecture**
16. **Model Selection**
    * **Objective:** Train and evaluate multiple regression models to determine the best-performing one for price optimization.
    * **Models Considered**:
      + Random Forest Regression
      + Ridge
      + Linear Regression
      + Lasso
      + Gradient Boosting
      + XGBoost
    * **Model Training**:
      + - Fit each model on the dataset using the model.fit() method.
17. **Prediction and Performance Evaluation**
    * **Objective:** Use trained models to predict optimal product prices and assess their accuracy.
    * **Performance Metrics:**
      + - **R² Score (R-Squared):** Measures how well the model explains variance in pricing.
        - **Mean Absolute Error (MAE):** Indicates the average absolute error in predicted prices.
        - **Root Mean Squared Error (RMSE):** Captures the standard deviation of residuals (prediction errors).
    * **Prediction Process:**
      + - Use model.predict() to generate price predictions for the test dataset.
        - Compute R², MAE, and RMSE using the actual vs. predicted values.
18. **Model Performance Comparison**
    * **Table Representation**: Performance metrics for each model will be compared in a structured table.
19. **Visualization of Model Performance**
    * **Bar Charts**: Compare the performance of different models based on R² Score, MAE, and RMSE.
    * **Color Gradient Representation**:
      + - **Green** → Better performance (higher R², lower MAE & RMSE)
        - **Red** → Worse performance (lower R², higher MAE & RMSE)
      1. **Model Explainability**
    * To ensure transparency in pricing predictions, we will use SHAP (SHapley Additive Explanations) and Permutation Importance to interpret the model’s decision-making process.
    1. **SHAP (SHapley Additive Explanations)**
    * **Objective**: Understand the contribution of each feature to the model’s predictions.
    * **Methodology**:
      + - Compute SHAP values for each feature.
        - Identify the most influential factors affecting product price predictions.
        - Visualize feature importance using summary plots, dependence plots, and force plots.
    * **Outcome:** Provides insights into how historical sales data, competitorpricing, and product attributes impact optimal pricing decisions.
    1. **Permutation Importance**
    * **Objective**: Assess feature importance by measuring the decrease in model performance when a feature’s values are randomly shuffled.
    * **Methodology**:
      + - Shuffle one feature’s values while keeping other features unchanged.
        - Measure the drop in R² score and increase in MAE/RMSE.
        - Rank features based on their impact on model performance.
    * **Outcome:** Identifies key pricing factors by evaluating how much each feature contributes to prediction accuracy.
    * **Visualization of Model Explainability**
      + **SHAP Summary Plot**: Highlights the overall impact of each feature on pricing predictions.
      + **Permutation Importance Bar Chart**: Ranks features based on their influence on model performance.
      1. **Visualization Flow Architecture**
20. **Pricing Trends**

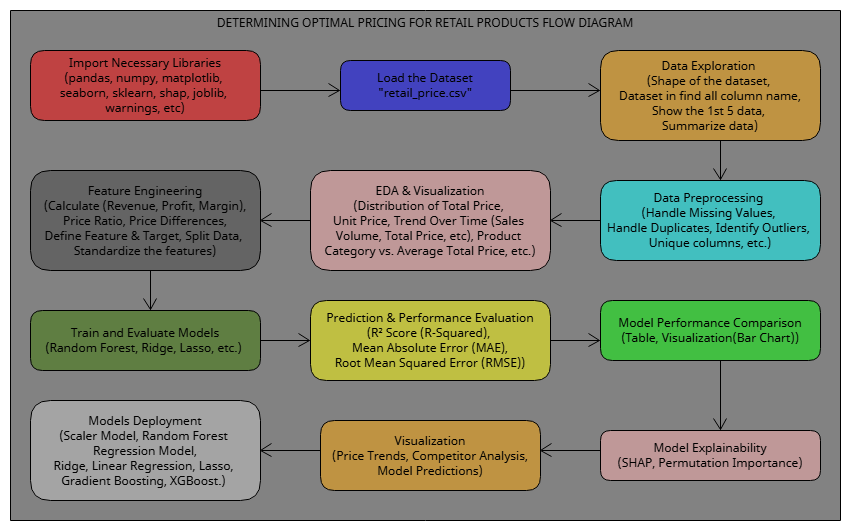
* **Objective:** Track the fluctuations in product prices over time, including competitor pricing and our own price trends.
* **Visualization:** A line plot comparing the monthly pricing trends of our product and three competitors.

1. **Competitor Analysis**

* **Objective:** Analyze the relationship between competitor pricing and total sales, as well as compare our product prices with competitor prices.
* **Visualization:** A scatter plot representation for:
  + Competitor Prices vs. Total Sales
  + Our Prices vs. Competitor Prices

1. **Model Predictions**

* **Objective:** Compare actual vs. predicted product prices for different regression models.
* **Visualization:** Scatter plots for each model’s predictions.
  + 1. **Model Deployment**
       - * Scaler Model
         * Random Forest Regressor Model
         * Rodge Model
         * Linear Regression Model
         * Lasso Model
         * Gradient Boosting Model
         * XGBoost Model
    2. **Complete Flow Diagram**

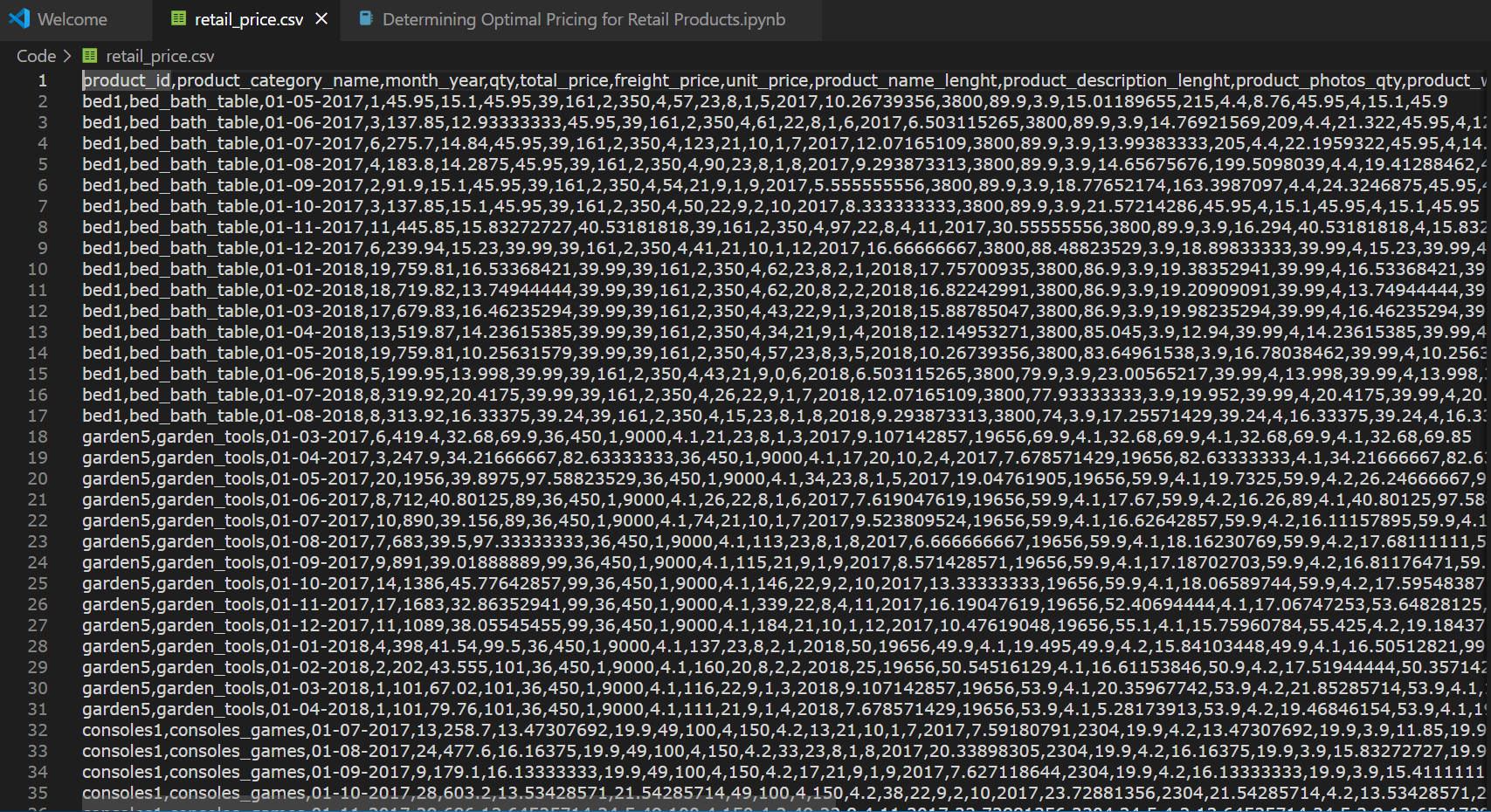


**Fig 1 - Complete Model Flow Diagram**

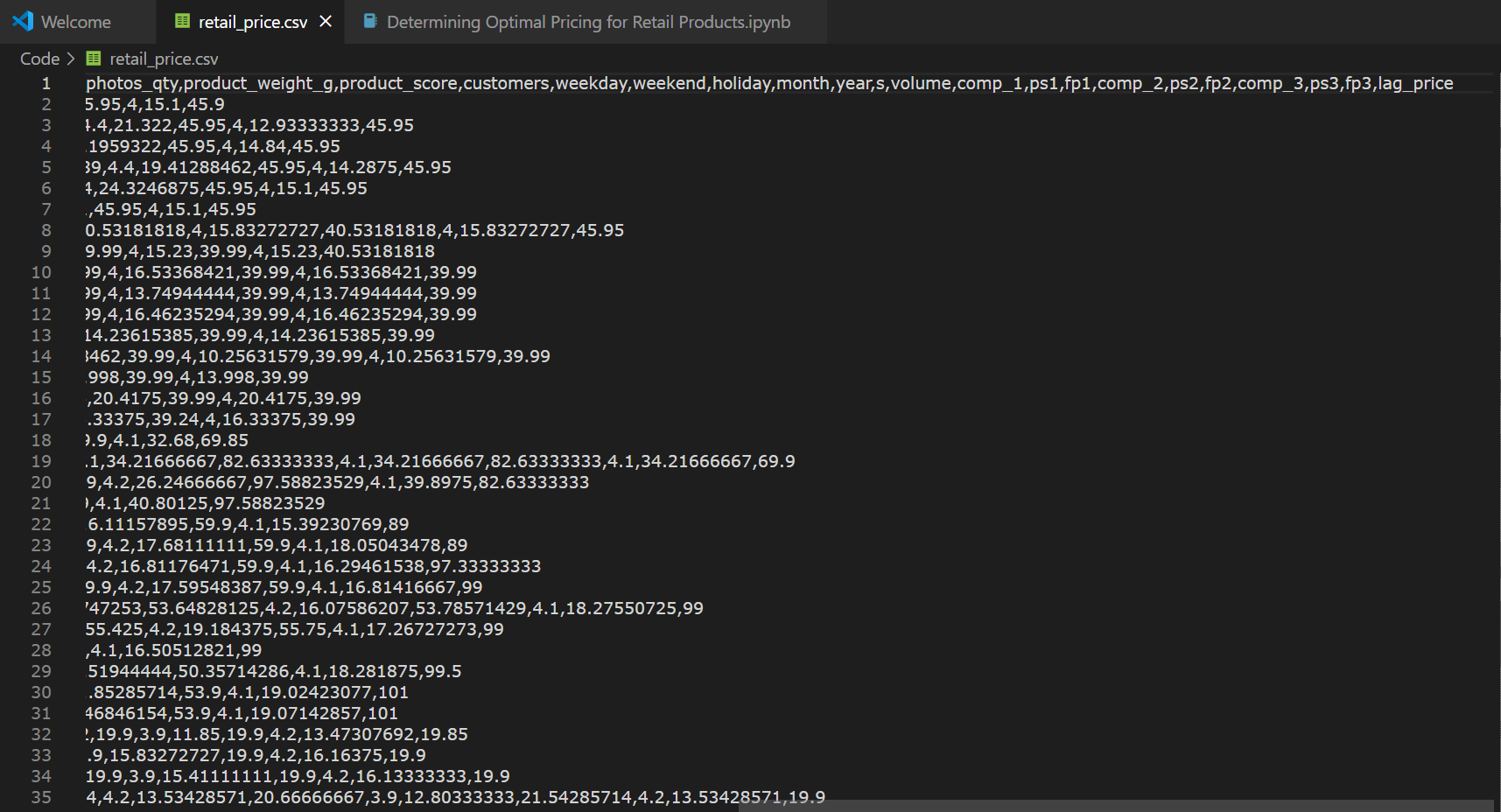
* + 1. **Implementation Workflow**
       1. **Import Libraries:** Load essential Python libraries for data manipulation, visualization, and model training, including Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, SHAP, and Joblib.
       2. **Load Dataset:** Read "**retail\_price.csv**" into a Pandas DataFrame (old\_df) and create a copy (df) for further processing.
       3. **Data Exploration**
          - Display dataset shape (df.shape), first 5 rows (df.head()), and column details (df.info()).
          - List all column names (df.columns).
       4. **Data Preprocessing:** 
          - Handle missing values by identifying columns with missing data and visualizing them using a heatmap.
          - Detect and remove duplicate rows.
          - Identify outliers in numerical columns using statistical analysis.
          - Separate categorical and numerical features.
          - Perform statistical operations on numerical columns.
       5. **Feature Engineering**
          - Compute additional features like revenue, profit, and price margins.
          - Calculate price ratios and differences across competitors.
          - Generate market demand indicators and temporal features (seasonality, month, and year effects).
          - Define input features (X) and target variable (y).
          - Split the dataset into training (80%) and testing (20%) using train\_test\_split().
          - Standardize numerical features using StandardScaler().
       6. **Model Training & Evaluation**
          - Train multiple regression models, including Random Forest, Ridge, Linear Regression, Lasso, Gradient Boosting, and XGBoost.
          - Make predictions using model.predict().
          - Evaluate performance using **R² Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE)**.
          - Compare model performance using a structured table and bar charts (green for better performance, red for worse).
       7. **Model Explainability**
          - Use **SHAP values** to identify key features influencing price predictions.
          - Apply **Permutation Importance** to measure the impact of each feature on model performance.
          - Visualize feature importance using SHAP summary plots and permutation importance bar charts.
       8. **Visualization Flow**
          - **Pricing Trends Over Time**: Line plot comparing price trends for our product vs. competitors.
          - **Competitor Analysis**: Scatter plot showing the relationship between competitor pricing and sales.
          - **Predicted vs. Actual Prices**: Scatter plots for different models comparing predictions to actual values.
       9. **Model Deployment**
          - Save trained models (joblib.dump(model, "model.pkl")).
          - Deploy the best-performing model for real-time price recommendations.
  1. **Dataset Details**

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| product\_id | Unique identifier for each product |
| product\_category\_name | Category of the product |
| month\_year | Date of the sales record |
| qty | Quantity of product sold |
| total\_price | Total price of the sale |
| freight\_price | Shipping cost |
| unit\_price | Price per unit of the product |
| product\_name\_lenght | Length of the product name |
| product\_description\_lenght | Length of the product description |
| product\_photos\_qty | Number of photos for the product |
| product\_weight\_g | Weight of the product in grams |
| product\_score | Rating or score of the product |
| customers | Number of customers who purchased the product |
| weekday | Number of weekdays in the month |
| weekend | Number of weekend days in the month |
| holiday | Number of holidays in the month |
| month | Month of the sale |
| year | Year of the sale |
| s | Likely a calculated field, possibly sales rate or similar metric |
| volume | Possibly the volume of the product |
| comp\_1 | Price of competitor 1's product |
| ps\_1 | Product score of competitor 1's product |
| fp\_1 | Freight price of competitor 1's product |
| comp\_2 | Price of competitor 2's product |
| ps2 | Product score of competitor 2's product |
| fp2 | Freight price of competitor 2's product |
| comp\_3 | Price of competitor 3's product |
| ps3 | Product score of competitor 3's product |
| fp3 | Freight price of competitor 3's product |
| (Target) lag\_price | Price of the product in the previous time period |

**Table 1 - Dataset Details**

****

**Fig 2 - Dataset**

****

**Fig 3 - Dataset**

1. **Implementation Details of Project**



<https://qr.me-qr.com/mobile/pdf/7eaefde9-239a-41e0-bef0-8ff4de6e6dd9>

**Fig 4 – Implementation Code PDF & Link**