PGP DSE

# FINAL REPORT

**Project Name: Dynamic Pricing through Historical Data**

1. Summary of problem statement, data and findings

Every good abstract describes succinctly what was intended at the outset, and summarizes findings and implications.

**Problem Statement:**

The project aimed to predict the total price a customer will pay for a product, given various features such as product price, quantity, discounts, category, and other time-related factors. While a mathematical formula could be used to calculate the price under fixed conditions, this project used machine learning to capture the more complex relationships and interactions that exist between features and the final price.

**Data:**

The data consisted of several features including:

Product details: Price, quantity, and discount

Temporal features: Year, month, day of the week

Category information: Product category such as meat, dairy, poultry

Customer-related features: Average spend per customer, revenue per product, etc.

**Findings:**

The Random Forest Regressor model was able to explain 99.999% of the variance in the TotalPrice, with very low prediction errors (MAE: 0.42, RMSE: 2.47, R²: 0.99998).

The most important features were DiscountedPrice, Quantity, and AvgSpendPerCustomer, as shown by the feature importance plot.

1. Overview of the final process

Briefly describe your problem solving methodology. Include information about the salient features of your data, data pre-processing steps, the algorithms you used, and how you combined techniques.

**Problem Solving Methodology:**

Data Pre-processing: Handled missing values, performed feature engineering (e.g., extracting year, month, day from SalesDate), and used one-hot encoding for categorical features like product categories.

Model Selection: The problem was approached using Random Forest Regressor to model the non-linear relationships in the data.

Evaluation: The model was evaluated using metrics like MAE, RMSE, and R². Cross-validation was used to ensure the model wasn’t overfitting.

Result: The model performed extremely well, predicting TotalPrice with high accuracy.

Techniques Combined:

Random Forest Regressor was used to capture complex relationships between features and the target.

Feature Engineering helped enhance the data’s predictive power.

Cross-validation was used to check the model's generalization ability.

1. Step-by-step walk through of the solution

Describe the steps you took to solve the problem. What did you find at each stage, and how did it inform the next steps? Build up to the final solution.

Step-by-Step Walkthrough of the Solution

**Data Exploration & Pre-processing:**

Removed irrelevant features (TransactionNumber, ProductName).

Extracted temporal features (year, month, day of the week) from SalesDate.

Applied one-hot encoding for categories and label encoding for Class.

Model Selection & Training:

Chose Random Forest Regressor for its ability to capture non-linear relationships.

Trained the model on the pre-processed data and fine-tuned it using cross-validation.

Model Evaluation:

Achieved high accuracy: MAE: 0.42, RMSE: 2.47, R²: 0.99998.

Feature importance analysis highlighted DiscountedPrice, Quantity, and AvgSpendPerCustomer as key predictors.

1. Model evaluation

Describe the final model (or ensemble) in detail. What was the objective, what parameters were prominent, and how did you evaluate the success of your models(s)? A convincing explanation of the robustness of your solution will go a long way to supporting your solution.

**Final Model: Random Forest Regressor**

Objective: Predict TotalPrice using features such as Price, Quantity, Discount, Category, and customer-related data.

Model Specifications: The Random Forest Regressor was implemented with multiple decision trees, using default settings for tree depth and number of estimators.

Performance Evaluation: The model achieved excellent accuracy. On the training set, it had minimal error (MAE: 0.165, RMSE: 0.97, R²: 0.999997). On the test set, the errors were slightly higher (MAE: 0.425, RMSE: 2.48), but accuracy remained exceptional (R²: 0.99998). The model explained 99.999% of the variance in TotalPrice.

Robustness & Feature Importance: Feature importance analysis confirmed that DiscountedPrice, Quantity, and AvgSpendPerCustomer were the most influential predictors. The results demonstrate the model’s strong predictive capabilities and reliability.

.

1. Comparison to benchmark

How does your final solution compare to the benchmark you laid out at the outset? Did

you improve on the benchmark? Why or why not?

**Benchmark & Model Performance**

Benchmark Goal: The objective was to develop an accurate model for predicting

TotalPrice, surpassing the accuracy of basic mathematical calculations.

Limitations of Simple Formulas: Traditional methods, such as multiplying Price ×

Quantity after applying discounts, fail to capture non-linear relationships and

complex feature interactions present in real-world data.

Model Superiority: The Random Forest Regressor effectively addressed these

limitations by leveraging multiple decision trees to model intricate patterns in the

dataset.

Performance Outcome: The model achieved exceptional accuracy (R²: 0.99998),

exceeding the benchmark expectations and demonstrating its effectiveness in

predicting TotalPrice with minimal error.

1. Visualization(s)

In addition to quantifying your model and the solution, please include all relevant visualizations that support the ideas/insights that you gleaned from the data.

The Feature Importance Plot highlights the most influential factors in predicting TotalPrice, with DiscountedPrice, Quantity, and AvgSpendPerCustomer ranking as the top contributors. The Learning Curve demonstrates how the model's performance improves as the training data increases, helping assess potential underfitting or overfitting. A well-balanced curve indicates that the model learns effectively without losing generalization ability. The Model Metrics Table provides a summary of key evaluation metrics, including MAE, RMSE, and R², which confirm the model's accuracy and robustness. Together, these visualizations validate the model’s effectiveness in capturing complex relationships and making reliable predictions.

1. Implications

How does your solution affect the problem in the domain or business? What recommendations would you make, and with what level of confidence?

**Business Impact**: The model can help businesses in dynamic pricing, where prices may vary depending on customer characteristics, quantity, discounts, and product categories.

**Recommendation**: Using the model, businesses can automate price prediction, ensuring more accurate and optimized pricing strategies.

Confidence: Given the model’s high R² and low error, the recommendations based on this model would have a high level of confidence.

1. Limitations

What are the limitations of your solution? Where does your model fall short in the real world? What can you do to enhance the solution?

**Model Limitations and Considerations**

The data dependence of the model is crucial, as its accuracy heavily relies on the quality and completeness of the dataset. Missing, inaccurate, or inconsistent data could negatively impact its predictive performance. While the model demonstrates strong performance, the risk of overfitting remains a consideration. It is essential to validate the model in real-world scenarios to ensure its robustness and generalizability beyond the training dataset. Additionally, the feature scope could be further expanded by incorporating additional relevant variables, such as customer behavior data, which may enhance the model’s predictive power and overall effectiveness.

1. Closing Reflections

What have you learned from the process? What would you do differently next time?

**Key Learnings**

This project highlighted the effectiveness of machine learning in predicting dynamic pricing, demonstrating how advanced algorithms can capture complex patterns beyond traditional mathematical calculations. Feature engineering proved to be a critical step, as extracting meaningful attributes from raw data significantly enhanced the model’s performance. Additionally, the project reinforced the importance of model evaluation and overfitting detection, ensuring that the model generalizes well to unseen data rather than just memorizing patterns from the training set.

**Future Improvements**

Incorporating additional features such as customer behavior trends or sales history could further enhance the model’s predictive capabilities. These additional data points could provide deeper insights into purchasing patterns, allowing for more refined and accurate pricing predictions.