Data Engineering Assignment Report

1. Task Overview

The analysis of supermarket transaction data encompassed two major analytical initiatives, each designed to address specific business challenges and opportunities within the retail operation.

The first initiative centered on developing a predictive framework for identifying high-performing supermarkets. This analysis went beyond simple revenue metrics to incorporate multiple performance indicators and their temporal patterns, enabling the identification of consistently successful stores. This knowledge is vital for understanding and replicating successful operational practices across the network.

The second initiative focused on understanding and optimizing voucher impact across different product categories. This analysis was crucial for maximizing the return on investment from the company's voucher program by identifying which product types showed the strongest positive response to voucher applications. The analysis required sophisticated data integration and modeling techniques to account for various factors that influence purchase behavior.

A major technical limitation arose when attempting to integrate the Promotions.csv dataset with other data sources. The resulting merged dataset became extremely large and exceeded our processing capabilities. This challenge forced us to exclude the promotions data from our final analysis, potentially limiting our understanding of promotional impacts on sales patterns. This limitation suggests that future analyses might benefit from.

2. Data Cleaning & Transformation

The data integration process began with a thorough assessment of the primary datasets:

The Sales dataset contained transactional data spanning two years, including critical metrics such as sales amounts, units sold, and customer information. This dataset required careful handling of temporal elements and customer identifiers to maintain analytical integrity.

The Items dataset provided product information including descriptions, types, brands, and sizes. This dataset served as a crucial reference for categorizing and analyzing product performance patterns. Particular attention was paid to standardizing product classifications and ensuring consistent naming conventions.

The Supermarkets dataset contained location information for all stores, which needed to be cleaned and standardized to ensure accurate geographical analysis. For the supermarket performance analysis, we developed a specialized data preparation approach. Transactional data was aggregated to create weekly performance metrics for each supermarket. This aggregation included calculating total revenue, unit sales, unique transaction counts, and voucher usage patterns. The weekly aggregation level was chosen as it provided a good balance between granularity and trend visibility. We engineered lagged features to capture temporal patterns in store performance. This included creating various lagged versions of key metrics such as previous week's sales, units sold, and transaction counts. These lagged features were crucial for capturing temporal dependencies and seasonal patterns in store performance. To handle missing values in the lagged features, we implemented a careful imputation strategy that considered the temporal nature of the data and the business context of each metric.

For the voucher analysis, we implemented a sophisticated data transformation pipeline. The process began by merging sales and items datasets using an inner join operation on the product code. This step was crucial for maintaining data integrity while ensuring that we had complete product information for all transactions. We then created a binary voucher usage indicator to clearly distinguish between transactions with and without voucher applications.

Temporal features were engineered to capture weekly and daily patterns in shopping behavior. This included creating day-of-week indicators and ensuring proper handling of holiday periods and special events that might influence shopping patterns.

For categorical variables such as product type, brand, and size, we implemented one-hot encoding to transform these attributes into a format suitable for machine learning models. This transformation was done carefully to avoid the dummy variable trap while preserving the essential categorical information.

3.Supervised Learning Models

3.1 High-Performing Supermarkets Prediction Model

The development of the high-performing supermarkets prediction model represented a more complex analytical challenge, requiring sophisticated handling of temporal patterns and multiple performance indicators.

3.1.2 Model Development and Implementation

The model was built using a **Random Forest** Regressor framework**,** enhanced with **GridSearchCV** for hyperparameter optimization. This approach was chosen for its ability to handle the complex relationships between various performance indicators while maintaining interpretability of feature importance.

The feature engineering process was extensive and included. Historical revenue metrics were carefully constructed to capture both absolute performance and growth patterns. Transaction count features were developed to represent both volume and efficiency of operations. Voucher usage patterns were incorporated to understand promotional effectiveness. Lagged indicators were created to capture temporal dependencies and seasonal patterns.

The hyperparameter optimization process explored various combinations of model parameters, ultimately identifying optimal settings:

The maximum depth of 10 provided the best balance between model complexity and generalization ability. The minimum samples split of 10 helped prevent overfitting while capturing meaningful patterns. The optimal number of estimators (300) provided stable predictions while maintaining computational efficiency.

3.1.3 Performance Analysis

The model achieved meaningful predictive power, with a cross-validation R^2 score of 0.531, indicating that it captured approximately 53% of the variance in store performance patterns. The test RMSE of 112.64 provided a concrete measure of prediction error in the same units as the target variable, allowing for practical interpretation of model accuracy.

The test R^2 score of 0.262, while lower than the cross-validation score, still indicated useful predictive power, especially considering the complex nature of retail performance prediction.

High Performers: Supermarkets **5, 16, 17, 23, 27** consistently exceeded the 125% revenue threshold.

Example: Supermarket 5's predicted revenue for week 26 was **404.90** (baseline mean: ~320).

Trend Analysis: These stores showed rising revenue trends over consecutive weeks, indicating sustained growth.

Recommendations:

- Inventory:- Stock high-demand items in top-performing stores. For instance, if Store 5 sells more electronics, ensure adequate stock.
- Staffing: Increase workforce during weekends to enhance customer experience.

These stores showed rising revenue trends over consecutive weeks, indicating sustained growth.

3.2 Voucher Impact Analysis Model

The voucher impact analysis model was developed with the specific goal of understanding and predicting how voucher applications affect units sold across different product categories. This understanding is crucial for optimizing voucher distribution strategies and maximizing their impact on sales.

3.2.1 Model Development and Implementation

The **Random Forest** Regressor was chosen as the primary modeling algorithm due to its ability to handle non-linear relationships and capture complex interactions between features. The model incorporated various feature types:

Voucher usage features were engineered to capture not just the presence of vouchers but also their historical effectiveness patterns. Product attributes were carefully encoded to maintain their predictive power while avoiding dimensionality issues. Temporal features were included to capture day-of-week effects and seasonal patterns.

The training process utilized an 80/20 split between training and testing data, with careful attention to maintaining the temporal ordering of transactions. This temporal consideration was crucial for ensuring that the model's predictions would be relevant for future voucher decisions.

3.2.2 Performance Analysis

The model's performance metrics revealed interesting patterns. The relatively low R^2 scores (Train: 0.041, Test: 0.032) initially might seem concerning, but deeper analysis showed that this was largely due to the inherent variability in customer responses to vouchers. This variability actually provided valuable insights into which scenarios showed more consistent voucher effectiveness.

Key Findings:

- Type 2 Items: Demonstrated the highest uplift (+0.18 units per transaction). If 1,000 transactions occur weekly, allocating vouchers to Type 2 items could yield +180 incremental units sold weekly.
- Types 1, 3, 4: Minimal uplift (+0.07 units), suggesting vouchers have limited impact on these categories.

Recommendations: Prioritize vouchers for Type 2 items to maximize ROI.

4. Business Insights

4.1 Supermarket Performance Patterns

The analysis successfully identified 71 consistently high-performing supermarkets for the set threshold of 25% more than the average of total stores sales consistently for 2 weeks, representing locations that maintained superior performance metrics across multiple dimensions. These stores demonstrated several common characteristics. They maintained remarkably stable weekly revenue patterns, suggesting effective inventory management and consistent customer service. Their transaction counts typically exceeded the network average.

4.2 Voucher Effectiveness Analysis

The voucher impact analysis revealed several nuanced patterns in customer response to voucher promotions. The effectiveness of vouchers showed significant variation across product categories, with some showing consistently strong positive responses while others demonstrated more variable results. The analysis identified specific product types where voucher promotions consistently drove increased units volume. This pattern recognition allows for more strategic voucher distribution, potentially increasing the overall return on promotional investments.