Group 6 IE4214 Report



Submitted by

Low Eeron A0216716E

Tan Yi Zhen A0223694Y

Muhammad Imtiyaaz A0216752E

Jiang Yuchen A0222949W

DEPARTMENT OF INDUSTRIAL SYSTEMS ENGINEERING AND MANAGEMENT

NATIONAL UNIVERSITY OF SINGAPORE

Semester 1 2023

**Part A**

**Introduction**

This part focuses on analyzing customer purchase probabilities from product assortments, utilizing data from 'assortment.txt' and 'probability.txt'. We employed a range of machine learning models, including Random Forest, Gradient Boosting, Support Vector Regression, Neural Network, and XGBoost, as well as the Multinomial Logit Model. Our goal was to identify the most effective model for predicting customer choices, assessed by comparing Root Mean Square Error (RMSE) across these models. The insights gained from this analysis are critical for understanding consumer behavior and enhancing retail decision-making strategies.

**Method 1: Predictive Modeling of Customer Purchase Probabilities**

Approach:

The methodology started with the extraction of data from 'assortment.txt' and 'probability.txt' using a custom load\_data function, converting the textual data into structured arrays suitable for machine learning. We then created a feature matrix (X) and a label matrix (y) to represent product assortments and their respective choice probabilities. For model implementation, we explored various machine learning models including Random Forest, Gradient Boosting, Support Vector Regression, Neural Network, and XGBoost. Each of these models was evaluated using Root Mean Square Error (RMSE) to measure the accuracy of their predictions.

Rationale:

Transforming raw data into an analyzable format is fundamental for any machine learning task. By encoding the presence of products in assortments as features, we enable the effective training of models to predict customer preferences. Diversifying the selection of machine learning models allows for a comprehensive comparison of different predictive approaches. The use of RMSE as the evaluation metric aids in quantifying the average magnitude of prediction errors, providing a clear understanding of each model’s predictive performance and helping to identify the most effective model for predicting purchase probabilities.

**Method 2: Multinomial Logit (MNL) Model**

Approach:

After loading the assortment data and the choice probability data, an array is created to contain the features of each product. 1 is added to the features of each product to include the intercept parameter. Next, the assortments and its choice probability is split into training data and testing data. Then, a function “log\_likelihood1” is created. This function uses the MNL formula and calculates the negative log likelihood of the training data. Then, the optimal coefficients are obtained by minimizing the negative log likelihood. After obtaining the optimal coefficients, the predicted choice probabilities of the testing data are determined using the optimal coefficients. The RMSE is calculated between the true choice probabilities and the predicted choice probabilities.

Results

Table 1: RMSE comparison between different models

|  |  |
| --- | --- |
| Model | RMSE |
| Random Forest | 0.0183 |
| Gradient Boosting | 0.0558 |
| Support Vector Regression | 0.1921 |
| Neural Network | 0.1148 |
| XGBoost | 0.0422 |
| MNL Model | 0.0744 |

The analysis revealed that the Random Forest model, known for its robustness and ability to handle complex datasets, outperformed other models with the lowest RMSE. This result validated our choice and demonstrated the model's effectiveness in predicting customer purchase probabilities within various assortments. Therefore the outcome of our result is the ‘Group6.txt’ text document file.

**Part B**

**Introduction**

In the dynamic field of market analytics, determining the most effective pricing strategy is crucial for maximizing revenue. This study delves into three distinct but complementary methodologies to identify the optimal pricing for 'product b', a key item in our dataset derived from 'assortment.txt' and 'probability.txt'. Each method - the Data-Driven Price Optimization using Polynomial Curve Fitting, the Multinomial Logit (MNL) Model, and the Price Impact Analysis and Revenue Optimization for Product B (PIARO-B) Method - approaches the problem from a unique angle, utilizing advanced statistical and machine learning techniques. By focusing on product b’s pricing impact and its selection probability in various assortments, our analysis aims to unveil the most profitable pricing strategy. The results of these methods offer valuable insights into the relationship between product pricing, consumer choice, and revenue generation, guiding strategic decisions in pricing policies.

**Method 1: Data-Driven Price Optimization using Polynomial Curve Fitting**

Approach:

We began our analysis by extracting data from 'assortment.txt' and 'probability.txt' using a custom function, parse\_array\_file, which converts data into structured lists. Our focus was on assortments where 'product b' is the sole product. For the curve fitting, we utilized the curve\_fit function from the scipy.optimize module to model the relationship between 'product b's price and its purchase probability. Lastly, in our revenue optimization step, we defined a revenue function to calculate expected revenue at different price points for 'product b', followed by a comprehensive search across a range of prices to find the optimal pricing strategy.

Rationale:

The focused data processing isolates the effect of 'product b's pricing on purchase probabilities, allowing for a clear understanding of its impact on customer behavior. By considering only assortments with 'product b', we eliminate confounding factors from other products. The use of polynomial curve fitting is ideal for capturing the complex, nonlinear relationships often found in pricing and demand scenarios. It helps in understanding how price variations of 'product b' influence customer purchase decisions. The revenue function ties the price of 'product b' directly to expected revenue, considering the modeled purchase probabilities. Exploring a wide range of prices ensures that we thoroughly identify the most effective pricing strategy for maximizing revenue.

Results

Through our analysis, we determined the optimal price for 'product b' to maximize revenue.

Our methodology has successfully identified a pricing strategy that is expected to maximize revenue from 'product b', providing a significant strategic advantage in the market. Optimal price: 3548 and Maximum expected revenue: 2373.705536155369.

**Method 2: Multinomial Logit (MNL) Model**

Approach:

After loading the assortment data and the choice probability data, an array is created to contain the features of each product. 1 is added to the features of each product to include the intercept parameter. Next, the assortments, where 'product b' is the only product, are being extracted out with its choice probability. Then, a function “log\_likelihood1” is created. This function uses the MNL formula and calculates the negative log likelihood of the filtered assortment. Then, the optimal coefficients are obtained by minimizing the negative log likelihood. After obtaining the optimal coefficients, Excel solver is used to determine the optimal revenue by making price as the changing variable.

Rationale:

By extracting out the assortments, where 'product b' is the only product, the optimal coefficients obtained are only affected by ‘product b’.

Results

By maximizing revenue, we determined the optimal price for 'product b'. Optimal price: 3351 and Maximum expected revenue: 2325.83. File name is ‘Group6 for part b.xlsx’ excel file.

**Method 3: Price Impact Analysis and Revenue Optimization for Product B (PIARO-B) Method**

Approach:

To determine the optimal pricing for product b from the provided "assortment.txt" and "probability.txt" data, a specific approach is employed. Initially, the script reads both files, with "assortment.txt" containing various product assortments and "probability.txt" detailing the corresponding choice probabilities for these assortments. The focus is narrowed to assortments that include only product b (products 6 to 10) or an outside option (product 0), to isolate the impact of product b’s price on its selection probability. For each relevant assortment, the script extracts and stores the choice probability of product b in `product\_b\_data`. A dictionary, `price\_levels\_b`, maps each product number to its respective price. The script then calculates the expected revenue for each instance of product b by multiplying its price with the choice probability. These calculations are compiled in `expected\_revenues`, a list of tuples containing the product number, price, choice probability, and the computed expected revenue. The entries in `expected\_revenues` are sorted by the revenue value in descending order, aiding in identifying the price level that maximizes revenue. This sorted list elucidates which price levels are most effective for product b in terms of revenue generation, based on the provided dataset. This approach provides a comprehensive analysis of the relationship between price levels and choice probabilities for product b, aiming to identify the optimal price point for maximum revenue within the given market conditions.

Rationale:

The rationale behind this approach is grounded in the principle of revenue maximization, a key objective in pricing strategy. By focusing on assortments containing only product b, we isolate the impact of its price on consumer choice, ensuring that our analysis is specific to this product without interference from other variables. This method leverages the detailed data from "assortment.txt" and "probability.txt" to understand how consumers respond to different price levels of product b in varied assortments. The calculated expected revenue at each price point serves as a direct measure of the financial outcome of pricing decisions. By sorting these revenues, we effectively rank the price points in order of their profitability, providing clear and actionable insights. This approach not only helps in identifying the optimal price that maximizes revenue but also aligns with practical business needs where understanding consumer preferences and their impact on revenue is crucial for making informed pricing decisions. It represents a data-driven methodology that combines statistical analysis with economic theory, offering a robust framework for pricing optimization in retail and e-commerce scenarios.

Results:

The results from the analysis present a consistent and compelling case for the optimal pricing of product b at $3000. The analysis repeatedly shows that this price point yields an expected revenue of approximately $2293.08.

**Part C**

**Introduction**

This part presents three methods for optimizing product assortments in retail to maximize revenue. First, the Structured Data-Driven Revenue Optimization Method analyzes product combinations using structured data. Next, the Multinomial Logit (MNL) Model offers statistical insights into customer preferences for specific product groups. Lastly, the Triple SKU Revenue Optimization (TSRO) Method identifies the most profitable trio of products. Together, these approaches provide a comprehensive toolkit for retailers to strategically select product assortments, enhancing profitability in a competitive market.

**Method 1: Structured Data-Driven Revenue Optimization Method**

Approach:

The methodology begins with data ingestion from 'assortment.txt' and 'probability.txt' using the ast.literal\_eval function, which translates text data into Python lists. This step is vital for organizing product assortments and their corresponding purchase probabilities into a structured format. In the next phase, a pricing dictionary is created, assigning specific prices to each product ID, which facilitates the direct association of products with their prices. We then develop a calculate\_revenue function to compute the expected revenue for each product assortment. For the revenue maximization analysis, the script iteratively examines each assortment-probability pair, focusing on assortments with three or fewer products, and calculates the expected revenue to identify the most profitable combinations.

Rationale:

The data ingestion and processing method is crucial for transforming raw data into a format suitable for analysis, ensuring accurate representation of product assortments. The establishment of a clear price mapping is integral for precise revenue calculations, linking each product to its monetary value. This approach allows us to quantify the total expected revenue from an assortment, which is key to revenue maximization analysis. The decision to focus on assortments with three or fewer products reflects a realistic approach to sales scenarios, simplifying the analysis. This targeted strategy is designed to reveal the most profitable product combinations, providing valuable insights for strategic assortment selection and maximizing revenue.

Results

Optimal Product Combination and Revenue:

The analysis identifies the optimal product combination as [0, 15, 8, 19], with an expected revenue of 2159.965 (excluding the placeholder '0' for no purchase).This combination suggests that products 15, 8, and 19 together offer the most profitable strategy within our constraints. Therefore the SKU is b,c,d with a revenue of 2159.

**Method 2:Multinomial Logit (MNL) Model**

Approach:

After loading the assortment data and the choice probability data, an array is created to contain the features of each product. 1 is added to the features of each product to include the intercept parameter. Next, the assortments that contain a subset of Products 0, 3, 9, 12, 19, 24 and 27, are being extracted out with its choice probability. Then, a function “log\_likelihood1” is created. This function uses the MNL formula and calculates the negative log likelihood of the filtered assortment. Then, the optimal coefficients are obtained by minimizing the negative log likelihood. After obtaining the optimal coefficients, a list is created to contain all the combinations of at most 3 products. The revenue for all the possible combinations is calculated and the combination that gives the maximum revenue is determined.

Rationale:

By extracting out the assortments that contain a subset of Products 0, 3, 9, 12, 19, 24 and 27, the optimal coefficients obtained are only affected by these products.

Results

The store should carry Products 3, 12 and 27, which brings a revenue of 2180. Therefore the SKU is a,c,f

**Method 3: Triple SKU Revenue Optimization (TSRO) Method**

Approach:

The script's approach involves processing and analyzing data to identify the most revenue-optimal assortment of SKUs in a retail context. Initially, it processes two key text files: 'assortment.txt' and 'probability.txt'. 'Assortment.txt' contains various combinations of product IDs representing different SKUs, while 'probability.txt' lists the corresponding choice probabilities for each assortment. The data from these files are parsed into lists and then used to form a DataFrame, `df\_combined`, where each row corresponds to a specific assortment and its associated choice probabilities. The script sets specific prices for six SKUs (a, b, c, d, e, f) and establishes a mapping from product numbers to these SKUs. The main analysis involves iterating over all possible combinations of exactly three SKUs. For each combination, the script calculates the average expected revenue based on the assortments that include only those SKUs. This calculation considers the probability of each product in the assortment being chosen and its price. The script then identifies the combination of three SKUs that yields the highest average expected revenue.

**Rationale:**

The rationale behind this script is grounded in the need for effective assortment planning and revenue maximization in retail management. By analyzing various combinations of SKUs and their associated choice probabilities, the script aims to uncover which combinations are most likely to attract consumer purchases and, consequently, generate the highest revenue. This analysis is crucial for retailers who need to make informed decisions about which products to stock together. The choice of focusing on combinations of exactly three SKUs is based on a specific business constraint or strategy, perhaps reflecting limited shelf space or a desire to simplify the product range. The approach of calculating average expected revenues for each combination allows for a comprehensive understanding of the potential performance of different assortments, guiding strategic decisions in inventory management and sales optimization.

**Results:**

The analysis conducted reveals that the optimal combination of SKUs to maximize revenue consists of SKUs b, c, and d. This specific assortment is projected to yield an expected revenue of approximately $1763.01.