*Submitted to the INFORMS 2023 BSS Data Challenge Competition*

FOR INTERNAL USE ONLY (leave blank)

Submission Number:

Team Number:

---------------------------------------------------------------------------------------------------------------------

Using a Hierarchical Model to Determine Optimal Price & Profit Forecasts for 2023 BSS Data Mining Challenge

Anubhav Shankar

[anubhavshankar1993@gmail.com](mailto:anubhavshankar1993@gmail.com)

**Abstract**: In the rapidly evolving world of eCommerce, strategic pricing plays a crucial role in driving profitability and competitive advantage. Accurate pricing influences consumer behavior and impacts a company's bottom line. In industries with fierce competition and dynamic pricing, such as eCommerce, finding the optimal price for a product becomes a complex yet essential task. This task will be tackled by creating a hierarchical/mixed-effect model to predict the optimal volume and then optimize using a suitable profit function to obtain the optimal price. The optimal price will then be used to forecast the profit for the desired lookahead window.

1. Introduction

The ‘unitsordered’ (renamed as ‘volume’ for the analysis) has a high level of variability, which the seller can control for. We can maximize the daily profit by accurately predicting the volume sold and optimizing for the price with a relevant profit function. To accurately predict the volume, we’ll make a hierarchical model controlling for the cost and individual ‘sku’ to capture the dynamic demand curves. The provided data will be split into training and validation sets, respectively, to create the model, and the model subsequently will be tested on a randomly generated test set having the same distributional assumptions as the original data to forecast the daily profit prediction for the ‘t+6’ lookahead window.

1. Methodology

A cursory glance through the abstract and the motivation sections might preordain using either a linear regression or a numerical method like extrapolation for the first objective and, consequently, use something like an Autoregressive Integrated Moving Average (ARIMA) for the second objective.

However, the methods above may fail to capture each product's inherent volatility and dynamic demand curves. So, a hierarchical/mixed effect model is more suitable for such a purpose. Additionally, a parametric model will be less computationally intensive and more intuitive.

1. Abbreviations and Acronyms

The following abbreviations will be used extensively throughout the document for brevity:

i.) ME -> Mixed Effect

ii.) ARIMA -> Autoregressive Integrated Moving Average

iii.) LME -> Linear Mixed-Effect

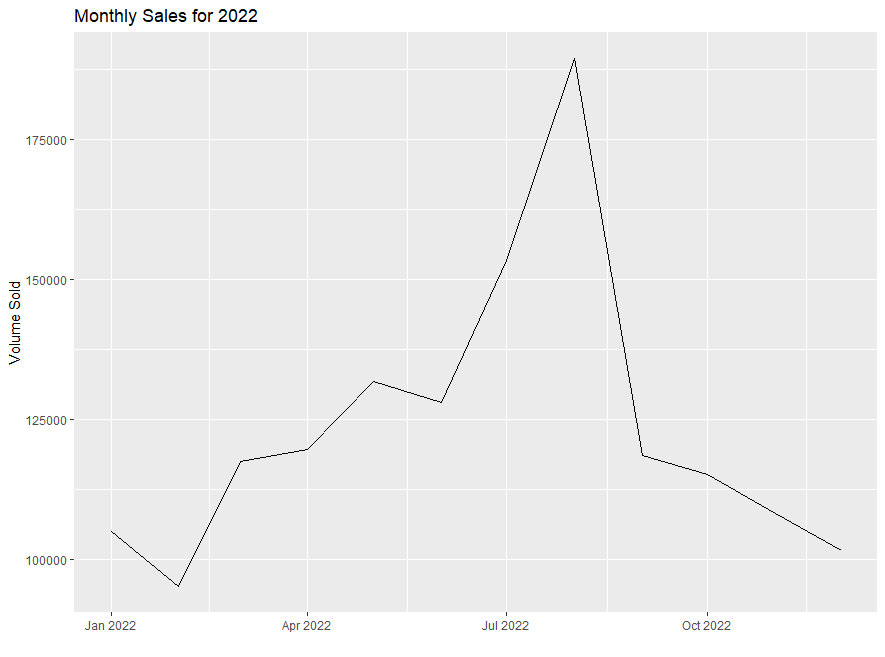
iv.) OLS -> Ordinary Least-Squares

v.) RE -> Random Effect (can be used interchangeably with (i) & (iii))

vi.) FFSKU47 -> File-Folders SKU 47

1. Initial Insights

Before we get into the application and understanding of ME models. Let’s briefly examine a few insights gleaned from the provided dataset. Let’s start by seeing the monthly sales and quarterly profit trends for 2022.

 A graph with a line going up

Description automatically generated

Figure 1: Monthly Sales trend for 2022 Figure2: Quarterly Profit trend for 2022

The figures above show that sales and profits fell sharply around the end of August. For more granularity, the volume distribution per day is shown below. The graph below shows the sales distribution is nearly normal, with Saturday as the worst-performing day. Such nuances will be useful to control for.

A graph with a red line

Description automatically generated A graph of blue bars with white text

Description automatically generated

Figure 3: Daily Sales Trend Figure 4: Sales per Product Category

Additionally, there are nearly 230 individual products that are sold. However, they can be grouped into nine (9) distinct types/categories, as shown in the graph above that tracks the sales performance of each category. ‘File Folders’ is the best-performing product type/category. Concurrently, we also know that each product can have multiple competitors. Let’s visualize the number of competitors per product category to get a bird’s eye view of what we can expect.

A graph of a bar chart

Description automatically generated with medium confidence Figure 5: Competitor count per product category

In addition to being the best-performing product category, ‘File Folders’ also has the highest number of competitors. While these overarching insights are helpful, let us dive deeper and see the demand curves for the ‘File Folders’ category and our assigned ‘sku’, ‘FFSKU 47’ for a snapshot in time.

A graph with red dots

Description automatically generated A graph of a product

Description automatically generated with medium confidence

Figure 6: Demand Curve for File Folders Figure 7: Demand Curve for FFSKU 47

Right away, we observe a nuance. While the demand curve for ‘File Folders’ as a category shows a nominal spread with a noticeable upward slope, the curve for FFSKU47 shows a higher spread and has a practically flat slope.

Additionally, as seen in Figure 8 below, we see another interesting phenomenon:

A graph with red dots

Description automatically generated Figure 8: Price vs. Revenue for FFSKU47

For the same spread, we observe an upward-sloping relationship between price and revenue, indicating that the revenue clocks an increase for a unit increase in the price!

1. Mixed Effect Models & Approach

Also known as Linear Mixed-Effect models, it is primarily used for modeling a mixture of fixed and random effects. Fixed effects refer to the typical main effects one would see in a linear regression model, i.e., the non-random part of a mixed model. In some contexts, they are referred to as the population average effect. Random effects are simply those specific to an observational unit, however defined. In our case, the observational units would be the products themselves.

The lmer() command, part of the ‘lme4’ package, can run an ME model in R [[1]](https://m-clark.github.io/mixed-models-with-R/random_intercepts.html).

We will create an ME model to predict the optimal volume and optimize it using the optim() command in the lme4 package to get an optimal price point.

Mathematically,

Volume = 1 + β1 \* X1 + β2 \* X2 + β3 \* X3 + β4 \* X4 + β5 \* X5 + β6 \* X6 + γ \* X7 + εi - (1)

In Eq(1), X1, X2, …… X7 -> The independent variables on which the dependent variable, ‘volume,’ will be modeled.

where,

εi -> The error/residual term

X1 -> comp\_1\_price\*

X2 -> comp\_2\_price\*

X3 -> cost

X4 -> Markup; coded as I(price-cost)

X5 -> salesdate

X6 -> weekday

X7 -> RE of cost for each ‘sku’; coded as (1 + cost | sku), the ‘1’ represents the intercept for the RE

The profit function conducts a row-wise operation by multiplying the predicted volume with the optimal price and subtracting the static costs. The ME model and the profit function are then fed into the optim() command within the defined bounds to arrive at the optimal price point.

\**Note: While the File Folders product category has five competitors, FFSKU47 only has two. Hence, only two competitor prices have been included in the model*.

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Estimate/Coeffs. | Std.Error | t-value |
|  |  |  |  |
| (Intercept) | 49.464\*\*\* | 1.772 | 27.915 |
| comp\_1\_price | -0.0032 | 0.003 | -1.004 |
| comp\_2\_price  cost | -0.0091\*\*\*  0.0343\*\*\* | 0.003  0.004 | -2.946  8.344 |
| I(price – cost) (Markup)  salesdate  weekdayMon  weekdayTue  weekdayWed  weekdayThu  weekdayFri  weekdaySat | -0.0208\*\*\*  -0.0025\*\*\*  0.0651  -0.0576  -0.0866\*  0.0262  0.0997\*\*  -0.0132 | 0.001  0.00009  0.049  0.05  0.086  0.0497  0.0486    0.0473 | -10.508  -27.893  1.328  -1.151  -1.731  0.528  2.050  -0.281 |
| Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |  |  |  |

1. Results

Table 1: Fixed Effects from the Hierarchical Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Groups | Name | Variance | Std. Deviation | Correlation |
|  |  |  |  |  |
| sku | Intercept | 10.9281 | 3.305 |  |
|  | Cost | 0.0011 | 0.034 | -0.85 |
| Residual |  | 6.2171 | 2.493 |  |

Table 2: Random Effects from the Hierarchical Model

1. Advantages of the Hierarchical Model
2. Less computationally intensive
3. Intuitive and relatively easy to replicate
4. The chosen variables lead to a relatively parsimonious model
5. Conclusion

From Table 1 (2-D) above, we can see that the fixed-effect coefficient for the markup is negative and is thus indicative of the fact that for our particular ‘sku,’ a higher price may, in fact, be more profitable. Also, given that ‘comp\_2\_price’ is statistically significant, it augurs well that the optimal price point predicted is near that price without exceeding it while also being within the ‘max\_price’ & ‘min\_price.’ Also, the calculated MAPE on the simulated test set was 4.52%, while it was 21.1% for the validation set.

1. Acknowledgements

This project has been a challenging endeavor. I want to thank Neal Fultz, Statistical Scientist, UCLA (nfultz@ucla.edu), for his valuable insights and fine-tuning my approach to the problem by familiarizing me with hierarchical models.

1. Gen-AI Disclaimer

Owing to the complex nature of the undertaking, LLMs like ChatGPT have been used extensively for debugging purposes and to understand the literature regarding the Hierarchical/ME models.

1. References

1. [Mixed Models with R](https://m-clark.github.io/mixed-models-with-R/random_intercepts.html)
2. [Intro to R Stats: Hierarchical Linear Models](https://methodenlehre.github.io/intro-to-rstats/hierarchical-linear-models.html#models-including-level-2-predictors)
3. Forecasting Principles & Practice [(a)](https://otexts.com/fpp3/useful-predictors.html) & [(b)](https://otexts.com/fpp2/gts.html)
4. [Hierarchical Linear Modeling (HLM): An Introduction to Cross-Sectional & Growth Modeling Frameworks](https://files.eric.ed.gov/fulltext/ED545279.pdf)
5. [lme4: Linear Mixed-Effects Models](https://rdrr.io/cran/lme4/)
6. GitHub

The codebase can be found in this [repository](https://github.com/aAnubhav2147/BSS-Competition).