E-commerce Product Optimization

Lauren Ally and Sean Merkle

Purpose

- To predict **key performance indicators** in the e-commerce space, specifically focusing on the core problem of **forecasting the number of units sold** at a given price.
- Leveraged **product-level data** that included price, historical sales, category, and title.

Background

- **~1.7 million** row dataset of Amazon US product listings. We **cleaned** the data, **encoded** categorical variables, and **processed** the title feature.
- Used **TF-IDF vectorization** to process the title feature, which focuses on **word importance**.
- Used **Sentence Transformers** to capture the **overall meaning of the title** regardless of exact wording.
- **Scaled** target variable due to **large variability** (boughtInLastMonth).

Methods

MLP - Python, Manual

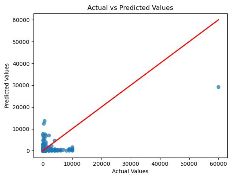
- **2 hidden layers** (e.g., 30 and 15 neurons)
- **ReLu** activation function
- **Gradient descent** & **L2** regularization to update weights
- **16,000** randomly sampled rows
- **Learning rate** = .01
- **10,000 epochs,** monitoring loss to prevent overfitting
- Analyzed **learning curves** and **residuals**, and **sensitivity analysis** to understand feature influence and validate the network's generalization capability

Bayesian Model - R, Packages

- Fit the model with **MCMC** using the **brms** package
- **Formula**: boughtInLastMonth ~ price + stars + reviews + isBestSeller + category
- warmup = **1000**, iter = **5000**, chains = **4**
- **5000** randomly sampled rows
- **Recasted** & **filtered categories** due to differences in train/test sets
- **Trace plots**, **posterior density plots**, and **posterior predictive checks** to interpret model behavior

```
Number of test observations before recasting: 1500
Number of test observations after recasting & filtering: 1494
```

MLP Results

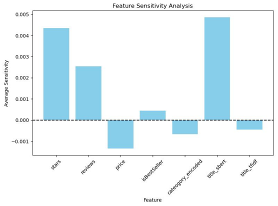


Performance Analysis

R2≈ **0.39**: Model explains less than half the variation in units sold, indicating a low to moderate degree of predictive power.
MSE: **1,021,124**. This number is quite large, and we believe it is largely **driven by outliers.**
RMSE ≈ **1010.51**. MAE ≈ **279.72**. Typical error magnitude — on average, predictions are off by ~280 units.

Coefficient Analysis

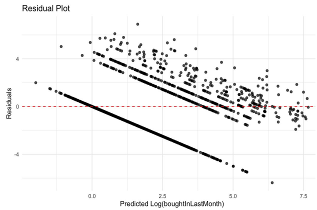
Stars: This feature had the **largest** coefficient (0.076), showing that this feature had a **significant positive** correlation with units sold. The next largest was **reviews** (0.025).
Title: Had **surprisingly little effect**, likely due to fluff in titles muddling the prediction. This highlights the **importance of more specific features** to capture the value of the item itself
Price: This feature had a **low negative value**, which directionally makes sense. However, it showed **less impact than we thought**. Moreover, we believe this is tied to the weak title analysis, leaving **price with a lower effect without the item context**.

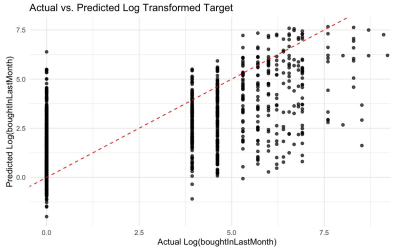


Bayesian Model Results

Performance Analysis

R2≈ **0.50**: Model explains about half the variation in units sold → **moderate degree of predictive power**
Error Metrics (on the log scale): imply the model errors by a **factor of about 3–5** on individual predictions (non-log space).
RMSE ≈ **1.98**, MAE ≈ **1.50**, Median Absolute Error ≈ **1.09**
Residual Analysis: The plot indicates the model may systematically **underestimate** sales for higher-selling products and/or **overestimate** at lower-selling ranges.





Coefficient Analysis

isBestSeller: Large positive coefficient (1.19) implies that bestsellers can expect an increase in units sold.
Category Effects: Certain categories (Health & Household, Kitchen & Dining) have high positive coefficients (4–7 on log scale), showing increases in expected units sold relative to the category.

Category-driven effects and **isBestSeller's** large coefficient show how product type + brand popularity can influence sales. We think **interaction terms** (price × category) or nonlinear methods can further refine predictions.

Conclusion

- Analysis revealed that **external factors and market dynamics** play a much larger role than anticipated.
- Models performed **moderately**, but **highlighted crucial areas for further research** and underscored the inherent **complexity of buying decisions** in the e-commerce environment.

Future Work

- **Predict optimal price** to complement predicted number of units sold depending on vendor preferences.
- We believe a big portion of why our model didn't perform as well as we had hoped was due to a **lack of product description**. Creating more **specific product columns derived from the title** may perform better than the NLP methods that we used.
- Create a UI in a **web application**. Ideally, we could use the product_id to **pull in an image** of the item in question, and include a slider to help vendors tailor their inventory and price.
- **Expand the feature set** by incorporating external factors. This would include things like **competitor pricing, seasonal trends, and market sentiment data** to refine price predictions and capture broader market dynamics.

References

Abigail, M. (2025, February 25). Optimizing E-commerce Supply Chains with Predictive Analytics [Article]. ResearchGate. https://www.researchgate.net/publication/389357716

Jakkula, A. R. (2023). Predictive Analytics in E-Commerce: Maximizing Business Outcomes. Journal of Marketing & Supply Chain Management, 2. https://doi.org/10.47363/JMSCM/2023(2)158