Do People Focused on Exercise Shop and Look the Same

John McElderry, Anvesh Ravipati, Austin Rose
University of Nebraska - Omaha

Introduction

The research project presented is to list a real-world problem, analyze the data presented, create models, and present the findings. For this project, Amazon purchase and customer data was given to identify a potential solution to a marketing problem. The problem presented in this presentation is how can Amazon increase sales of exercise equipment. This question can be broken down into parts.

First, the data will be leveraged to identify what groups of people are buying fitness related products. Within that subgroup of the dataset, a time series model will be leveraged to answer what is the trend of those types of purchases over time. With that information, predictions can be made about future revenue or what targets need more focused advertising campaigns.

This project also wants to highlight if those customers who buy exercise related products are also buying other products that are closely related to exercise or even if they are buying some products unrelated to exercise. This can also help with sales as Amazon can now use purchases both inside and outside of the fitness realm to suggest future purchases.

This project should either challenge or confirm current expectations for this type of analysis. The authors current expectations include:

- Individuals who buy health/fitness related items will be clustered as younger customers than jump a few age ranges then peak again at an older age grouping.
- Sales counts will increase around the new year then tail off. This is more of a reflection of a "New Year's Resolution" bump.

Research Question

The core question this project hopes to answer is:

- Can one decipher what individuals who buy exercise related items might typically look like, and can this information be used to predict future purchases?

By answering this question, this project research will aim to assist Amazon in increasing sales for the exercise and fitness market.

Data Source

The dataset used, spanning 2018 – 2022, contains purchase data from the online retail giant, Amazon. Roughly 5,000 Amazon customers based in the United States were paid to provide their purchase history and demographic information during the set time periods. All participants were 18 years or older and English speakers. The data collection and publication were approved by the MIT Institutional Review Board.

Methodology

Data Preparation

The dataset initially contained a wide range of product categories from Amazon, spanning various domains. The goal was to isolate data specific to health and fitness products.

To identify relevant categories, a list of key terms was created and used to filter the dataset to retain only those entries with categories matching these the selected terms. The purchase categories discovered using those keys terms were:

Hand strengthener	Muscle roller	Exercise mat
Weight rack	Weightlifting belt	Exercise band
Exercise strap	Exercise block	Strength training machine
Abdominal exerciser	Wearable weight	Fitness bench
Exercise machine	Fitness stepper	Fitness hoop
Exercise step platform	Weight plate	Fitness equipment

Initially, the project aimed to include both health and fitness categories. However, as the analysis progressed, it was discovered that focusing on fitness categories provided more meaningful insights for customer profiling as the definition of "health" regarding merchandise quickly became blurred. As a result, the scope was narrowed to only include fitness categories, ensuring a more targeted and impactful analysis. Focusing more on fitness made the category choosing process to be more concrete. Health is a broad topic and can be challenging to define. Health can be exercise and healthy eating, but it can also be hygiene, sleep, supplements, physical endurance, mental health, and many other categories presented in the dataset. An example to show this would be does one consider yoga pants to be an item focused on health and exercise or simply a different type of pants. The major problem was that choosing a definition of health was far more difficult than simply choosing a more defined and concrete category to analyze, such as exercise.

After solidifying what categories would be used, the purchases with the categories related to exercise were isolated. Additionally, the purchase data was marked in a binary fashion for easier processing and use in the decision tree model. The date columns were also updated to the proper format to be utilized by the time-series model, Prophet. A similar approach was applied to other non-exercise categories in the dataset by giving a binary code to those who had bought an item in a category before and those who had not. The categories in this projects scope were:

Has purchased motherboard	Has purchased nutritional supplement	Has purchased planner
Has purchased downloadable video game	Has purchased games	Has purchased protein supplement powder
Has purchased pet food	Has purchased food storage container	Has purchased ABIS book
Has purchased golf bag	Has purchased skin cleaning agent	Has purchased instrument parts and accessories
Has purchased outdoor living	Has purchased Amazon book reader	Has purchased dietary supplements
Has purchased skin moisturizer	Has purchased massager	Has purchased headphones
Has purchased toys and games	Has purchased herbal supplement	Has purchased medication

Has purchased cookie	Has purchased wearable computer	Has purchased ABIS sports
Has purchased coffee	Has purchased weigh scale	Has purchased physical video game software
Has purchased sugar candy	Has purchased mineral supplement	Has purchased vegetable
Has purchased meat	Has purchased pet supplies	Has purchased health food
Has purchased physical movie	Has purchased speakers	Has purchased snack chip and crisp
Has purchased skateboard	Has purchased protein drink	Has purchased vitamin
Has purchased swimwear	Has purchased sporting goods	Has purchased sports drink

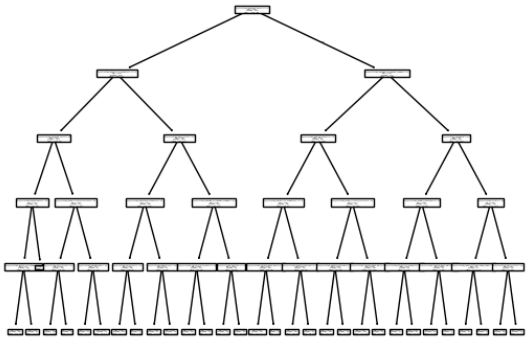
This method allowed for a comparison of what products customers are interested in based on whether they are interested in exercise or not. Only dates before March 2023 were used as the data begins to trail off after that point.

Using the cleaned purchase data, the customer data was filtered to include only customers who had made a fitness purchase. This data was used to create summary graphs to frame the dataset. Both Python and R were used in cleaning the data.

Data Model

Decision Tree

The decision tree model was chosen to specifically look at what characteristics of customers who bought exercise-related products had associated with themselves. The first challenge was to decide which customers bought exercise equipment and which customer did not to avoid skewing the results towards the larger share of the data (non-exercise customers). The solution was to take the average amount of money each person in the data set spent on exercise and look at those who spent more or less than the overall average. Once that issue was solved, a decision tree model was able to be applied.

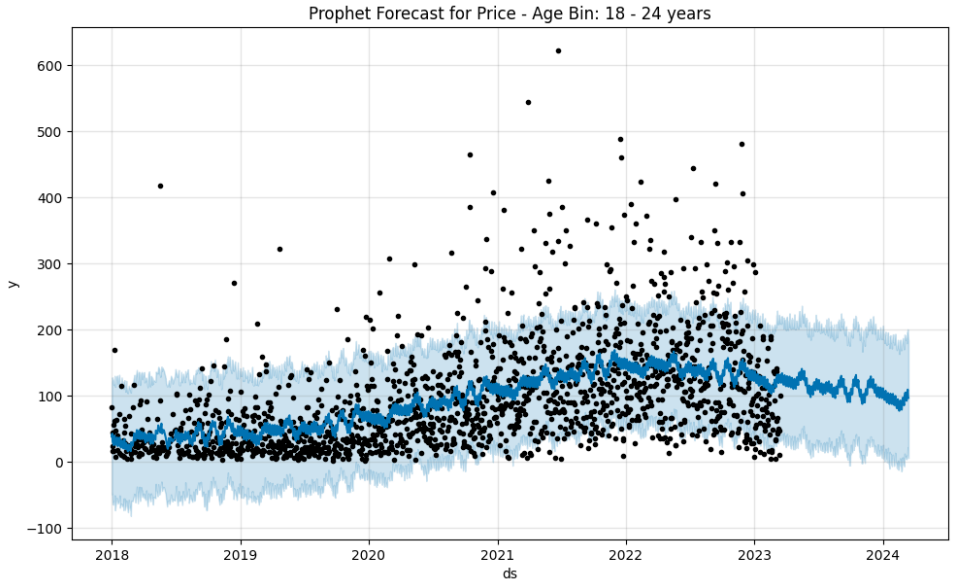


Python was used to create the decision tree. The “sklearn” library was the tool of choice for this model. The tree had a max depth of 5 and produced an 80.6% accuracy result. The survey results did not have meaningful results in splitting our decision tree, so customer purchase history was looked at. We looked to see if they had bought a certain category of item in the past and if that had any dependence on if they were regular fitness buyers.

Prophet

The Prophet model was applied to analyze trends in the data based on the quantity and number of purchases for a given day (price and number of purchases for a given day were summed up). The analysis was conducted at multiple levels:

1. **Overall Data**:
    a. Captured the general trends and seasonality patterns across all purchases.
2. **Age Bins**:
    a. Analyzed the data segmented by different age groups to identify distinct patterns or preferences among various age demographics.
3. **Race-Based Analysis**:
    a. Explored trends specific to different racial groups to understand variations in purchasing behavior.



The Prophet library in Python was used for this analysis. No major code edits were made. The results were displayed using the matplotlib.pyplot library.

Results

Prophet model revealed that the trends across all groups—overall data, age bins, and racial categories—were similar. The key observations were:

1. **Gradual Increase in 2021**:
    a. Purchase quantities and numbers showed a steady rise starting in 2021.
2. **Peak in 2021-2022**:
    a. The trends reached their highest point between late 2021 and early 2022, indicating a significant surge in purchasing activity during this period.
3. **Slight Decline Thereafter**:
    a. Following the peak in purchases, there was a gradual decline in purchasing trends, suggesting stability or a reduction in demand.

Decision-Tree model revealed that there are certain categories of amazon products that those interested in fitness are also interested in. With this information, there is a deeper understanding of what categories of items on Amazon to recommend to those who buy fitness products. It is important to push the products that people tend to buy the most along with fitness products as these may get more sales. However, recommending categories of products that a few people interested in fitness also buy is important as well because this could lead to more purchases further removed from exercise and then they will be directed to new areas of products that will increase revenue.

- **Highest priority recommendations for fitness-interested buyers:**
    - Sporting goods, protein supplement powder, nutritional supplements, outdoor living, weight scale, swimwear, and food containers
- **Other recommendations**
    - Herbal supplements, vitamins, outdoor living, and skin moisturizers

Limitations

Most limitations with this project come from a lack of background data and data volume. Although the goal is to try and identify a strategy to increase sales revenue, it is difficult to suggest alternatives without understanding current marketing and sales strategies. The focus of this project also limits the dataset to a small subset that could lead to misleading results. Additionally, this data has sample limitations as the sample demographics do not accurately reflect the broader U.S. population. This may lead to a bias in the results.

Conclusion

Through a decision tree analysis, a deeper understanding of the types of products favored by individuals interested in exercise was obtained. This analysis aims to recommend products to exercise enthusiasts more effectively. Popular Amazon categories were tested, revealing which ones are commonly purchased by these individuals. Expected categories included sporting goods, protein powder supplements, nutritional supplements, weight scales, swimwear, and food containers—all related to health, dieting, and fitness.

Unexpectedly, other categories such as herbal supplements, vitamins, outdoor living, and skin moisturizers also showed significant interest among exercise enthusiasts. Including these broader categories can enhance the recommendation algorithm's accuracy and expand the scope of product suggestions for those interested in exercise, offering a more comprehensive shopping experience.

The time-series analysis suggested that most demographics behave the same when it comes to fitness purchase timing. The largest knowledge gain from this model is what demographics purchase more fitness equipment.

This project would suggest that it is best to target younger customers who also purchase sporting goods, protein supplement powder, nutritional supplements, outdoor living, weight scale, swimwear, and food containers.

What We Learned

At least 88% of the work on this project was simply cleaning and arranging the data to better understand what we were specifically looking at. This helps frame both the research question and approach. I would also add that the project gave me more experience with GitHub which I both enjoyed (as it seems to be a straightforward tool) and appreciated the practice Googling how heck to do the things I wanted. Data file sizes are the bane of my existence *-Austin*

This project felt like an application oriented to me. It has been a year since I joined the master's program, so I was always learning but doing this project wanted to use all my skills. One of the highlights for me was being part of a highly functional and collaborative team. It felt like working with real co-workers, where everyone contributed equally and played their role effectively. This was a refreshing experience compared to previous group projects, where the workload often felt uneven. This project helped improve my teamwork skills significantly, and I truly appreciated the positive dynamic within the group. *-Anvesh*

From this project it seemed as if we had very little data to work with and that extracting any meaningful information would be very difficult and somewhat impossible. After diving deeper into the data set, we had to look for information about customers in a different more unique way rather than the survey only. While the survey had some useful information, we found that looking at purchase history was where the most valuable information was. Extracting this information and turning it into something that was plottable and sufficient for producing a model was one of the most difficult tasks. The reason it was difficult was not necessarily the logic behind extracting the information but the method to do so. Figuring if we should look at if someone had bought an item in the past was sufficient to say if someone is “interested” in a category or if we had to fig deeper and see if they had spent a certain dollar amount compared to others was a better method. In summary, the most difficult portion of the project that was a huge learning point for me was that building models that are meaningful is always possible if you have meaningful data. Being given meaningful data is rare, however, cleaning and preparing data correctly and effectively can take a dataset that seems meaningless and difficult to use and transform it into a data set that can be used to extract important information and build models that can lead to large and accurate decisions. *-John*