SENG 550: Analyzing the Market Viability of Amazon Products for E-Commerce Entrepreneurship

Group Members:

Name	UCID
Benson Li	30091566
Jack Li	30096387

Introduction and Motivation

E-commerce has reached US$6.31 trillion worth of sales as of January 2023 (Forbes, 2023). Simultaneously, dropshipping is an industry with a US$243 billion valuation with an expected annual growth rate of 24.39% (Oberlo, 2023). With Amazon accounting for 37.8% of e-commerce sales (Forbes, 2023), our objective is to conduct a comprehensive analysis of Amazon products with the aim of identifying potential products for dropshipping. Our analysis will look into trends using data such as: product categories, customer ratings, number of reviews, and price ranges. Ultimately, our goal is to find promising products or product types for sale and establish viable prices.

Data Collection

For this objective, the dataset "Amazon Products Dataset 2023 (1.4M Products)", sourced from: https://www.kaggle.com/datasets/asaniczka/amazon-products-dataset-2023-1-4m-products/, will be used. This dataset contains two .cvs files, one being 'amazon_products.csv' and the other titled 'amazon_categories.csv'. amazon_categories.csv contains a list of all Amazon product categories with a designated category ID while amazon_products.csv contains the products with useful information such as the Amazon Standard Identification Number(ASIN), product name, star rating, number of reviews, price, category ID, and if the product has had the "Best Seller" label.

Data Inspection and Validation

The initial step for inspecting and validating the Amazon Products Dataset was to remove any product entries that contained null data in any field. Only 142 entries contained these errors, leaving 1,426,196 entries with non-null data. Duplicate entries were also checked for, however none were found and deletion was not necessary. For category IDs, some were found to be invalid. These consist of IDs that fit within the range of category IDs, but do not exist in the amazon_categories dataset, thus products that do not contain a matching category ID that is

valid to be joined between the two dataset, leaving 1,388,159 complete and valid product entries. However, while these items cannot be analyzed within their categories, they are still useful data entries containing a product, therefore we have decided to keep these entries within the larger dataset for analysis of all possible products.

Data Filtering (extracting subsets of data based on select features or feature values)

Considering the products need to be analyzed from a holistic view, minimal data filtering is required. However, some necessary subsets were still extracted. As a simple subset, previous or current best seller products were selected; the total number of these products is 8520. On the other hand, for the majority of cases where the best seller label is not used, the label was excluded. Furthermore, some analysis will be conducted on the product categories separately; there are 248 such categories. Also, metadata that was not useful for executing operations such as the image and product URLs were also excluded from the dataset.

Data Transformations

In the original dataset, the price of the item was denoted by two values, 'price', the current price of the item, and 'listPrice', the original pricing of the item without discounts. For items without any such discounts, the original price of the item was placed in the 'price' value, with the 'listPrice' being 0. In the case of these items, we duplicated the 'price' value into the 'listPrice' field so that all items could be analyzed by their original pricing within a single data field.

Other data transformations include:

- Changing the .cvs format of the entry separated by commas to instead be separated by a pipe delimiter. This was done in the cases where the product name has commas, these commas could possibly interfere with an operation. It was also done for visual clarity for humans.
- Joining the two .csv files, amazon_products and amazon_categories, so the category displays the name of the category instead of the number.
- Creating two classes of key-value pairs, one for all products containing the RDDs of product ASIN keys, and one for analyzing categories with a category ID key.

Exploratory Data Analysis

For our Exploratory Data Analysis(EDA), we had initial plans for examining certain data trends while also branching out to investigate other trends from revelations which we had during the process. These included:

- Finding the highest rated products overall. However, this data was not useful because the number of perfectly scored 5 star items were too many and many of the perfectly scored items had zero written reviews, making the validity of the rating hard to substantiate.
- Finding the products with the most amount of reviews. This allowed us to see the products that customers were most compelled to write reviews for while also being an indicator for how well the product sold.
- Identifying the categories with the highest number of products in them. This data is useful in identifying the popularity of products to sell among businesses, but could also be viewed as an indicator for market saturation.
- Calculating the average rating of items in each category. These can identify the customer satisfaction towards certain products and generally how well these products perform against other product categories. In particular, gift cards are by far the highest rated group of products which can indicate the safety of venturing into that particular category.
- Calculating the average price of items in each category. This shows an estimate of what an item is priced around for each category for properly pricing a product and shows a possibility of undercutting other products.
- Finding the highest rated product in the top 20 categories by average rating. This shows a spread of the top items in each top category and can be used to identify common trends beyond categories.
- Finding the top 5 products based on the score of rating multiplied by reviews divided by the price. This should determine the supposedly best products based on the number of reviews at the best cost.

Visual representations of each trend found during the EDA can be viewed in the Jupyter notebook.

Model Building and Results

During the model building, we decided on attempting to predict how well an item has sold based on the price and rating. In order to do this, we started off by creating a regression dataset that had the number of reviews, price and rating. In order to prevent data skew, we removed data that had 0 reviews. We then used LabeledPoint to assist in the regression process.In order to maintain a standardized range, we had to utilize feature scaling, which involved using the mean and standard deviation of the feature values. The mean was subtracted from the values and the result of that was divided by the standard deviation. Next, the dataset was split into the training and validation sets using the specified weights of .8 for training and .2 for validation, where we had 236638 points in the training set and 59161 in the validation set for a total of 295799 points. Following that, we worked on creating and evaluating our baseline model, which involved finding the average number of reviews of 870 and calculating the root mean squared error for our model. Moving on, we began training our regression model using MLlib, we used LinearRegessionWIthSGD to get a linear regression model and to predict a sample point. From that point, we tried a different model, using a 5th degree order model to see if there was

improvement in the prediction and we repeated the steps for creating and training the model using the training set. Lastly, we plotted the predicted number of reviews for each of our models and the expected model.

This model aims to predict potential future Amazon products to be drop shipped based on early pricing and rating data before a market niche is established or for wholesaling in an already mature market.

References

[1] https://www.forbes.com/advisor/business/ecommerce-statistics
[2] https://www.oberlo.com/statistics/dropshipping-market