Market Research Analytics of Toy Products on Amazon

A Data Analytics Group Report

by

Chinmay Bake, Holly Johns, Rey Lopez

QMST 5336

Fall 2019

Introduction

Mr. Big is a business owner with a background in sales and marketing. He wishes to expand his business ventures by exploring the possibility of starting an online toy selling business. Before he begins, he wants to initiate some market research and requests our assistance. Specifically, Mr. Big has asked our team to analyze patterns within a historical dataset of Amazon toys sales and to offer him information to help him make his decision. Mr. Big's initial budget for this project is $2,000.

Problem Definition

The goal of this analytical report is to analyze a historical toy sales database and strategize ways Mr. Big might maximize the profitability of a toy-selling business he is considering pursuing. To solve this problem, we implemented a three-tier approach. First, we analyzed the historical dataset and compiled a descriptive analysis that would reveal patterns within. Next, we continued the analysis of the dataset using predictive statistics to help us establish a relationship between certain variables. Finally, we utilized prescriptive statistics to help us identify ways to maximize profitability. Specifically, through this historical analysis, we wish to answer the following three questions: What does the dataset reveal that would interest Mr. Big and his online business? Are there any patterns within this historical dataset that can help predict whether certain categories of toys will sell better than others? What recommendations can we make to help Mr. Big maximize his profitability?

Data Resource

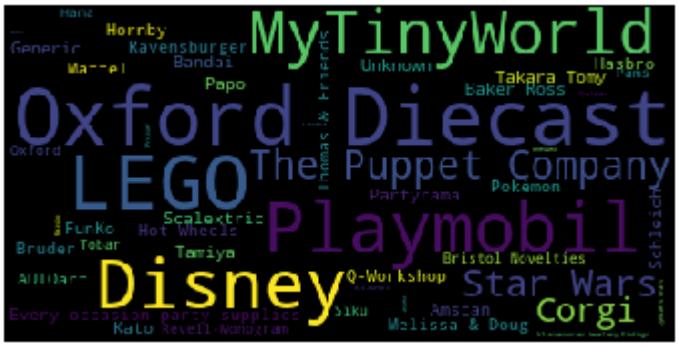
For this project, an open source dataset called 'Amazon Toys' was selected. This dataset was downloaded from Kaggle (www.kaggle.com), an online platform that connects a worldwide community of data scientists and machine learners with a variety of datasets. Founded in 2010 and acquired by Google in 2017, Kaggle provides a wealth of data set for exploration, analysis and competitions. Additionally, this data contains over 10,000 historical toy sales records from Amazon, a popular, ecommerce retailer. As such, we are confident the data downloaded for this project are reliable, credible, and appropriate for this project.

In this dataset, there are 10K records, 17 columns, and a wealth of information. For example, one column includes a list of companies that have sold the items and the average price for which it sells for. Another column includes detailed comments from customers who have purchased each item. To further analyze the data, we filtered the data three ways. First, we removed outliers. For example, one product had a seller's price listed for $17 but the dataset showed one seller sold the item for over $1900. Because this number was so drastically different that other sales for the same item, we excluded this item from certain parts of analysis in our data. Next, to improve the reliability of this data, we deleted records that were missing data from certainly fields. For example, we excluded records with no historical sales information. Similarly, during certain parts of our analysis, we only included toys that had a rating on 4.7 stars or better.

Descriptive Analysis

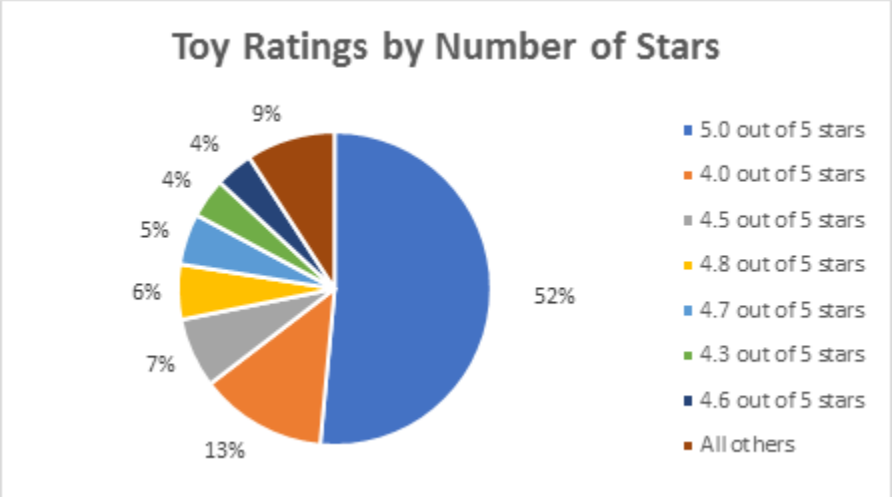
Using descriptive statistics, we created visuals to show what the dataset revealed, visuals that would be of interest to Mr. Big as he explores his idea of starting an online toy selling business. Through analysis, we determined that some manufacturers are more popular than others. To show Mr. Big which companies historically manufacture the most popular toys sold, we analyzed a column that contained *manufacturer* information. In reviewing this column, we only included manufactures with toys that had received ratings of 4.7 stars or better. To best show this information visually, we created a Word-Cloud with a list of the top 50 manufacturers (Figure 1).

Figure 1: Top 50 Manufacturers



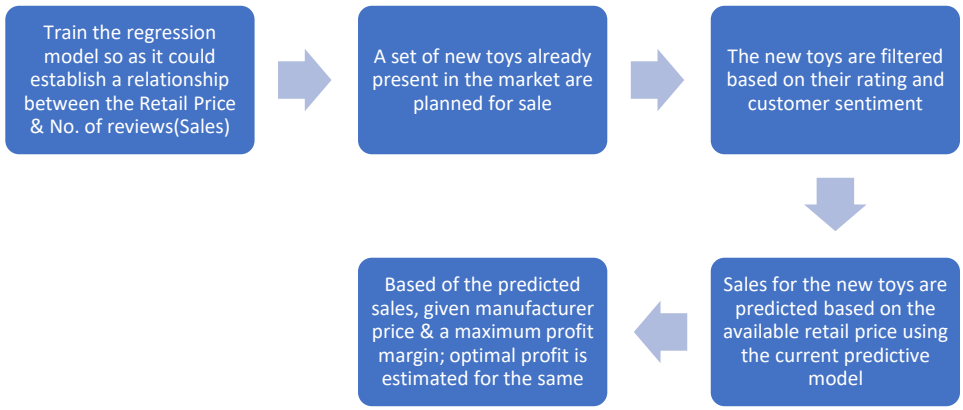
Further, we reviewed the '*amazon-category-and-sub-category*' column to identify the top categories that sell the most toys and concluded that the top category of toy sales are Hobbies and Die-Cast & Toy vehicles, and Figures & Playsets. Next, we looked more closely at the '*average-review-rating'* column to determine how many products received the highest ratings based on the number of stars they received. After reviewing the data in this column, we decided to combine all records that received 4.5 or less into one category called *other,* as combined, this group only accounted for 9% of the data. From this data, we also determined that over 52% of all toys sold received a rating of 5 out of 5 stars (Figure 2). T

Figure 2: Ratings

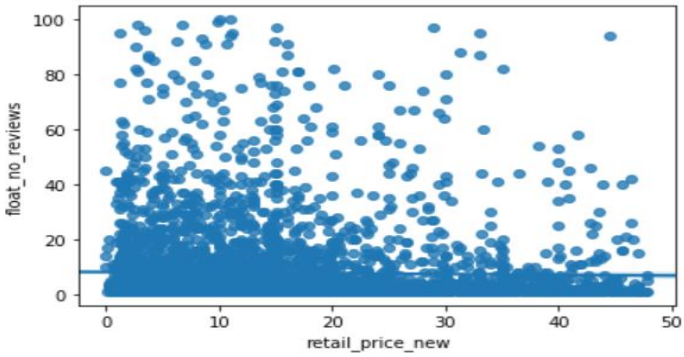


Decision Making

A Generic Workflow of Our Decision Framework



Training the Regression Model



1) In the above figure, we have a scatter plot having **number of reviews** as float_no_reviews on the y axis and **retail price** as retail_price__new.
2) Our further analysis is based on a presumption that **number of reviews** for each toy product would be **synonymous** to its respective **sales numbers**.
3) The figure above illustrates a relationship between the **sales** and the **retail price** which has been extracted from the dataset.
4) Clearly, the relationship does not showcase any linear trend despite us getting rid of the outliers. Hence, we need to utilize a methodology which could **nonlinearly** map the values on y axis against the values on x axis.
5) We utilize **RandomForestRegressor** library in Python in order to fit the values.

Evaluating the Model Performance

MODEL PARAMETERS	MODEL 1	MODEL 2	MODEL 3
Random State	42	0	42
Estimators	1000	2000	2000
MODEL OUTCOMES			
Mean Absolute Error	9.3551	9.342	9.352
Mean Squared Error	260.15	259.87	260.06
Root Mean Squared Error	16.12	16.12	16.125

1) We can observe that manipulating the number of decision trees in form of estimators does not impact RMSE significantly.
2) The RMSE of **16.12** is itself strikingly close to the **average sales number**.
3) We proceed with **Model2** for making further predictions on the model.

Sentiment Analysis

1) We now have a set of toys which are **already available** in the market; but we now have a decision to make on which ones to sell. **These toys are a subset plucked from our historical dataset.** Once we narrow down on them, we could forecast the optimal profit which could be made from selling them.
2) The first phase of filtering these toys is by analyzing the **customer comments** they have against them. These toys have been a part of our historical dataset and we have got hold of the customer comments for each of them.
3) Each toy has multiple comments populated against it from multiple customers and we need to see a bigger picture out of it to gauge an **overall sentiment** of the customers.
4) Thereby, all these comments are preprocessed, amalgamated and sent through the **VaderSentiment library** in Python for each toy. The Sentiment analyzer quantifies these comments for each product in a dictionary and provides a score on a scale of 1.
5) The scale is distributed based on a **Positive, Negative & a Neutral factor** of the outcome of the Sentiment Analyzer.
6) Based on a preliminary analysis of the mean positive, negative & neutral values, we set a **threshold value of 0.25 or greater** for a product to be classified as a product with **Positive sentiment**, **0.15 or greater** for a product to be classified as a product with **Negative sentiment** and rest of the products are tagged as products with **Average Sentiment.**

Finally, we assort certain set of toys (13 to be precise) based on criteria as Sentiment **Positive** only and Rating greater than or equal to **4.7**

	product_name	Sentiment	New_rating	retail_price_new	price_new
4845	Trading Card Sleeves - 50 Ultra Pro Standard S...	Positive	5.0	12.95	2.79
8467	Vauxhall Viva AYR Burgh Police Unit Beat Car P...	Positive	5.0	9.99	23.39
3922	OMYGOD PINK BUTTERFLY PARTY MASK	Positive	4.7	13.47	1.99
7264	Brookite Cutter No.2 Kite	Positive	5.0	14.99	9.00
7968	My Doll 16.5" rag doll with tartan dress	Positive	5.0	17.99	34.75
4180	Richmond Toys Motormax 4.5-inch London Series ...	Positive	4.8	8.72	6.99
6200	Happy 40th Birthday Jumbo Banner Pink & Silver	Positive	4.7	1.81	2.20

Predict Sales Based on the Existing Retail Price

1) The new toys we are planning to sell, have a **Manufacturer Suggested Retail Price** (MSRP)included with them from the dataset.
2) We could **predict the sales** (no. of reviews) for these toys based on our existing **Non-Linear Regression model.**
3) The model predicts sales values with these retail prices as illustrated in the below table;

PREDICTED SALES	RETAIL PRICE	MANUFACTURER PRICE
6.80973622	12.95	2.79
8.45234474	9.99	23.39
5.40601667	13.47	1.99
7.57510614	14.99	9
4.49898488	17.99	34.75
16.67350952	8.72	6.99
3.76819417	1.81	2.2
2.33434264	14.93	14.94
17.09180476	12.42	40
1.481	8.09	29.21
7.05756443	6.8	7.99
11.39164777	1.95	7.99
3.05412583	16.99	3.67

Forecast Optimal Profit

1) From the above subset of data, we could frame an optimization problem to determine the optimal profit from selling above toys.
2) Profit earned from each toy could be computed by the **product of Sales & the amount earned out of each toy.** Retail Price would be the price for which we essentially plan to sell the toy and Manufacturer Price would be the price for which we would purchase the toy for. Thereby in our optimization problem, we would want to be maximizing below objective function for computing optimal profit;

$$Profit = \sum(Sales \times (Retail\ Price - Manufacturer\ Price))$$

3) The set of **decision variables** in our problem would the **retail prices** for each toy.
4) In the next stage, we phrase constraints for the optimization problem.
5) The maximum amount which we would be investing is **2000** and we would also define a profit margin for the retail price as below;

We set a maximum **30% percent profit margin of our retail price** by having the retail price estimated **1.42** times than the manufacturer price. In a nutshell, our constraints are;

1) $\sum Manufacturer\ Price\ Sale \leq 2000$
2) $Retail\ Price \leq 1.42 \times Manufacturer\ Price$

Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
C2	RETAIL PRICE	3.9618	6.80973622	6.80973622	1E+30	6.80973622
C3	RETAIL PRICE	33.2138	8.45234474	8.45234474	1E+30	8.45234474
C4	RETAIL PRICE	2.8258	5.40601667	5.40601667	1E+30	5.40601667
C5	RETAIL PRICE	12.78	7.57510614	7.57510614	1E+30	7.57510614
C6	RETAIL PRICE	49.345	4.49898488	4.49898488	1E+30	4.49898488
C7	RETAIL PRICE	9.9258	16.67350952	16.67350952	1E+30	16.67350952
C8	RETAIL PRICE	3.124	3.76819417	3.76819417	1E+30	3.76819417
C9	RETAIL PRICE	21.2148	2.33434264	2.33434264	1E+30	2.33434264
C10	RETAIL PRICE	56.8	17.09180476	17.09180476	1E+30	17.09180476
C11	RETAIL PRICE	41.4782	1.481	1.481	1E+30	1.481
C12	RETAIL PRICE	11.3458	7.05756443	7.05756443	1E+30	7.05756443
C13	RETAIL PRICE	11.3458	11.39164777	11.39164777	1E+30	11.39164777
C14	RETAIL PRICE	5.2114	3.05412583	3.05412583	1E+30	3.05412583

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
B15	OPTIMAL PROFIT MANUFACTURER PRICE	628.8391813	0	2000	1E+30	1371.160819

Upon running the solver solution, we determine that an **optimal profit of $629** could be earned by having set certain values for Retail Price as indicated in the Final Value column in the above Sensitivity Report.

Analytics & Recommendations

1) We recommend that this analysis be utilized for products with an estimated retail price of **less than $50** because the baseline for this decision-making process was capped at $50 while preproccesing the data for analysis.
2) We recommend the sentiment analyzer be coupled with average rating of a toy product while shortlisting products based on their market rating.
3) A **shadow price of 0** from the solver solution clearly implies that changes in the maximum amount of investment could be varied and that essentially won't make a substantial difference on the optimal profit.

Critical Thinking

1) As mentioned earlier while analyzing the regression model, the Root Mean Square Error of the predicted value is very close to the average of the actual Sales number (the one which is predicted). This is implying that the predictions are not highly accurate with this model.
2) This research is based only on historical data. Startup costs and continuing expenses such as shipping fees and packaging costs were not factored into this analysis yet are essential to this equation.
3) The assumption that No. of Reviews would be synonymous to Sales impacts the accuracy of the predictive model.
4) Accuracy of the sentiment analyzer cannot be properly gauged.
5) Since the accuracy of the regression model is not very high, the subsequent relationship which we draw out of the regression model is not very concrete.

Appendix A:

Attached code of the demonstrated analysis.

QMSTProjectUpdatedCode.ipynb

Appendix B:

Attached solution for optimization solver.

Optimization Solver.xlsx