

Unlocking the Secrets of Amazon's Bestsellers: Insights and Strategies

Advanced Data Analytics Project

Team "Vector":

Ariadnee Ziady,
Christina Kuleshevich,
Gabriela Garcia Gonzalez,
Rebecca Yoder,
Vishwajit Mohan Kumar



University of Colorado **Boulder**

Agenda

- Introduction
- Exploratory Insights
- Modeling
 - Logistic and Linear Regressions
 - Tree Regression
 - Tree Regression with Bagging
- Key Results
- Business Recommendations

Introduction

- Ecommerce is a growing industry (total expected revenue to hit US\$5.5 trillion by 2027)
- Amazon has market capacity of US\$2.13 trillion
- We chose to use product sales data from Amazon
- We aim to answer two questions:
 - Can we predict which items in the data will be best sellers based on the given information?
 - Can we create a model that predicts the volume of each product that has been purchased?

201. (n.d.). *eCommerce Sales & Size Forecast*. Retrieved October 4, 2025, from <https://www.trade.gov/e-commerce-sales-size-forecast>

CompaniesMarketCap.com. (March 27, 2025). Leading e-commerce companies worldwide as of March 2025, by market cap (in billion U.S. dollars) [Graph]. In Statista. Retrieved October 04, 2025, from <https://www-statista-com.colorado.idm.oclc.org/statistics/245340/leading-large-cap-e-commerce-companies-market-cap/>

Data: 42,675 observations across 17 variables

	title	rating	number_of_reviews	bought_in_last_month	current.discounted_price	price_on_variant	listed_price	is_best_seller	is_sponsored
1	BOYA BOYALINK 2 Wireless Lavalier Microphone for iPhone ...	4.6 out of 5 stars	375	300+ bought in past month	89.68	basic variant price: 2.4GHz	\$159.00	No Badge	Sponsored
2	LISEN USB C to Lightning Cable, 240W 4 in 1 Charging Cabl...	4.3 out of 5 stars	2,457	6K+ bought in past month	9.99	basic variant price: nan	\$15.99	No Badge	Sponsored
3	DJI Mic 2 (2 TX + 1 RX + Charging Case), Wireless Lavalier ...	4.6 out of 5 stars	3,044	2K+ bought in past month	314.00	basic variant price: nan	\$349.00	No Badge	Sponsored
4	Apple AirPods Pro 2 Wireless Earbuds, Active Noise Cancell...	4.6 out of 5 stars	35,882	10K+ bought in past month		basic variant price: \$162.24	No Discount	Best Seller	Organic
5	Apple AirTag 4 Pack. Keep Track of and find Your Keys, Walle...	4.8 out of 5 stars	28,988	10K+ bought in past month		basic variant price: \$72.74	No Discount	No Badge	Organic
6	Texas Instruments TI-84 Plus CE Color Graphing Calculator, ...	4.6 out of 5 stars	44,522	100K+ bought in past month		basic variant price: \$99.95	No Discount	Best Seller	Organic
7	Apple AirPods 4 Wireless Earbuds, Bluetooth Headphones, P...	4.5 out of 5 stars	13,466	10K+ bought in past month		basic variant price: \$88.11	No Discount	No Badge	Organic
8	Apple AirTag. Keep Track of and find Your Keys, Wallet, Lugg...	4.6 out of 5 stars	38,105	10K+ bought in past month		basic variant price: \$23.04	No Discount	Best Seller	Organic
9	Complete Protect: One plan covers all eligible past and futur...	4.0 out of 5 stars	4,380	N/A	16.99	basic variant price: nan	No Discount	No Badge	Organic
10	Apple iPad 11-inch: A16 chip, 11-inch Model, Liquid Retina ...	4.7 out of 5 stars	7,308	10K+ bought in past month		basic variant price: \$284.05	No Discount	Best Seller	Organic
	is_couponed	buy_box_availability	delivery_details	sustainability_badges	image_url	product_url	collected_at	product_category	
	Save 15% with coupon	Add to cart	Delivery Mon, Sep 1	Carbon impact	https://m.media-amazon.com/images/I/71pAqIVe3L_AC_U...	/sspa/click?ie=UTF8&spc=MTo4NzEzNDY2NTQ5NDYxNDQ...	2025-08-21 11:14:29	Phones	
	No Coupon	Add to cart	Delivery Fri, Aug 29		https://m.media-amazon.com/images/I/61nbF6aVIPL_AC_U...	/sspa/click?ie=UTF8&spc=MTo4NzEzNDY2NTQ5NDYxNDQ...	2025-08-21 11:14:29	Laptops	
	No Coupon	Add to cart	Delivery Mon, Sep 1		https://m.media-amazon.com/images/I/61h78MEXoJL_AC_...	/sspa/click?ie=UTF8&spc=MTo4NzEzNDY2NTQ5NDYxNDQ...	2025-08-21 11:14:29	Laptops	
	No Coupon				https://m.media-amazon.com/images/I/615UJ2aKoEL_AC_U...	/Apple-Cancellation-Transparency-Personalized-High-Fidelit...	2025-08-21 11:14:29	Phones	
	No Coupon				https://m.media-amazon.com/images/I/61bMNCeAUAL_AC_...	/Apple-MX542LL-A-AirTag-Pack/dp/B0D54ZTHY/ref=sr_1_5...	2025-08-21 11:14:29	Phones	
	No Coupon				https://m.media-amazon.com/images/I/71AyPChhFdL_AC_...	/Texas-Instruments-TI-84-Graphing-Calculator/dp/B00TFY...	2025-08-21 11:14:29	Other Electroni	
	No Coupon				https://m.media-amazon.com/images/I/61iBtxCUabl_AC_U...	/Apple-Bluetooth-Headphones-Personalized-Effortless/dp/...	2025-08-21 11:14:29	Phones	
	No Coupon				https://m.media-amazon.com/images/I/71rP7f78eFL_AC_UL...	/Apple-MX532LL-A-AirTag/dp/B0CWXNS552/ref=sr_1_8?dib...	2025-08-21 11:14:29	Phones	
	Save 50% with coupon				https://m.media-amazon.com/images/I/71tZ0n2xP9L_AC_U...	/Asurion-Complete-Protect/dp/B07RZ3LSHM/ref=sr_1_9?di...	2025-08-21 11:14:29	Other Electroni	
	No Coupon			Energy efficiency	https://m.media-amazon.com/images/I/61aPY8odPSL_AC_...	/Apple-iPad-11-inch-Display-All-Day/dp/B0DZ75TN5F/ref=...	2025-08-21 11:14:29	Cameras	

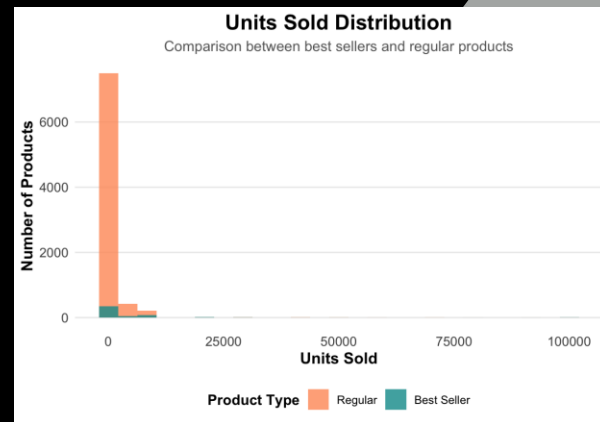
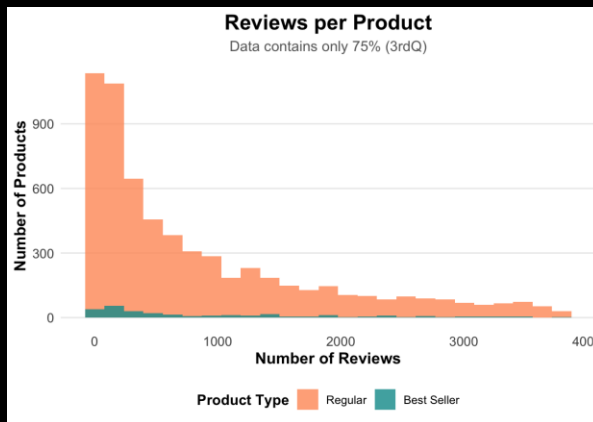
Objective

"To analyze the powerful factors that determine the success of Amazon's bestselling products, and how to leverage insights to maximize sales and optimize strategies".

Exploratory Insights

1. Right-skewed distributions

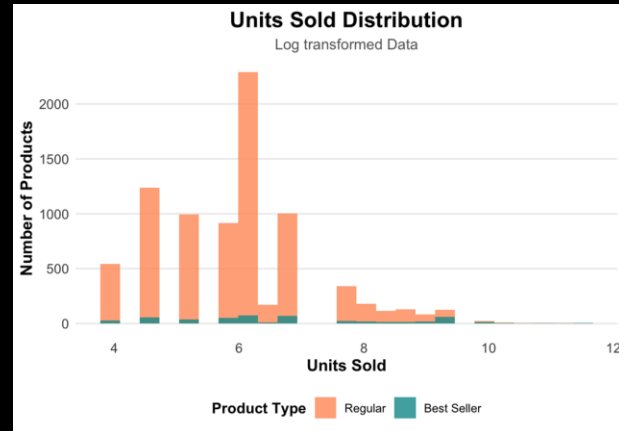
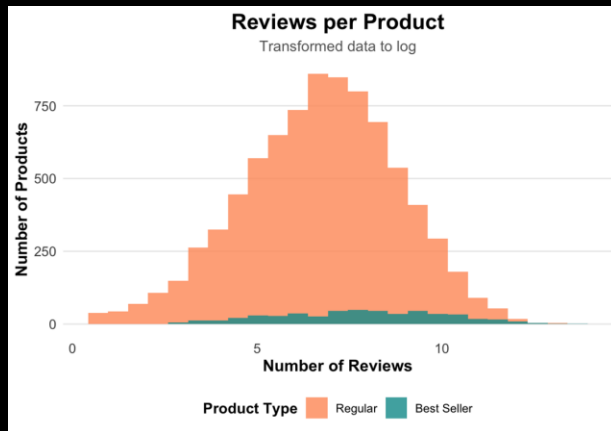
A few products dominate in Quantity Sold and reviews.



Exploratory Insights

1. Right-skewed distributions

Log-transformed plots show clear separation between bestsellers and others.



Exploratory Insights

2. Price “sweet spot”

Bestsellers cluster in **mid-range prices**: not the cheapest, not premium.



Exploratory Insights

3. Reviews & Sales

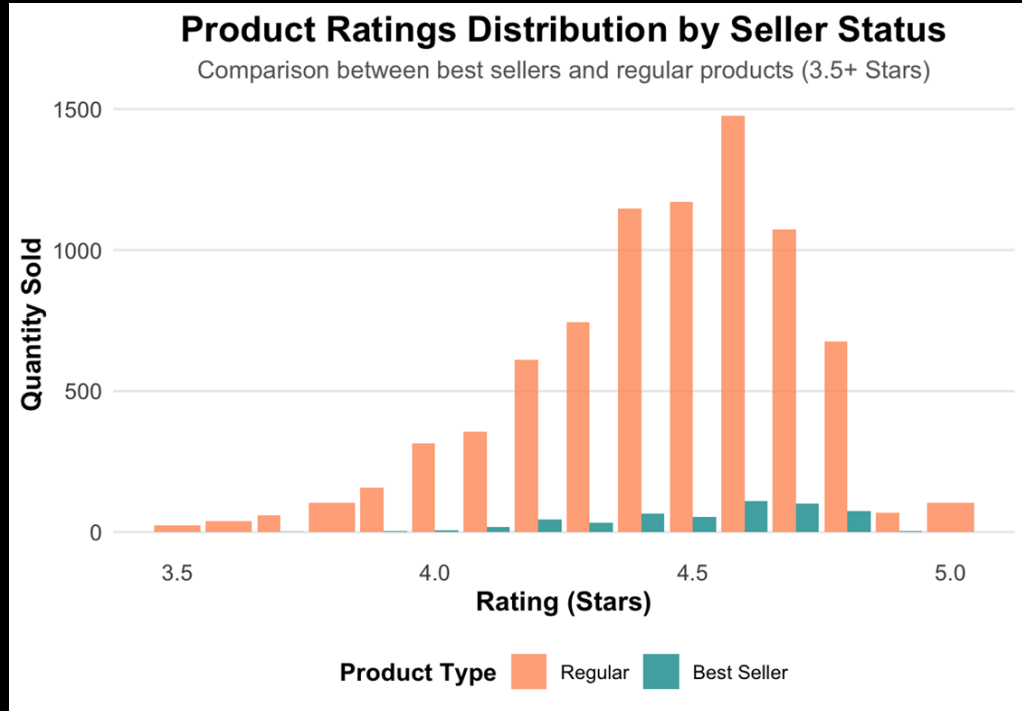
Products with higher reviews: higher likelihood of being bestsellers.



Exploratory Insights

4. Ratings

High ratings are common across all products: not a strong differentiator alone.

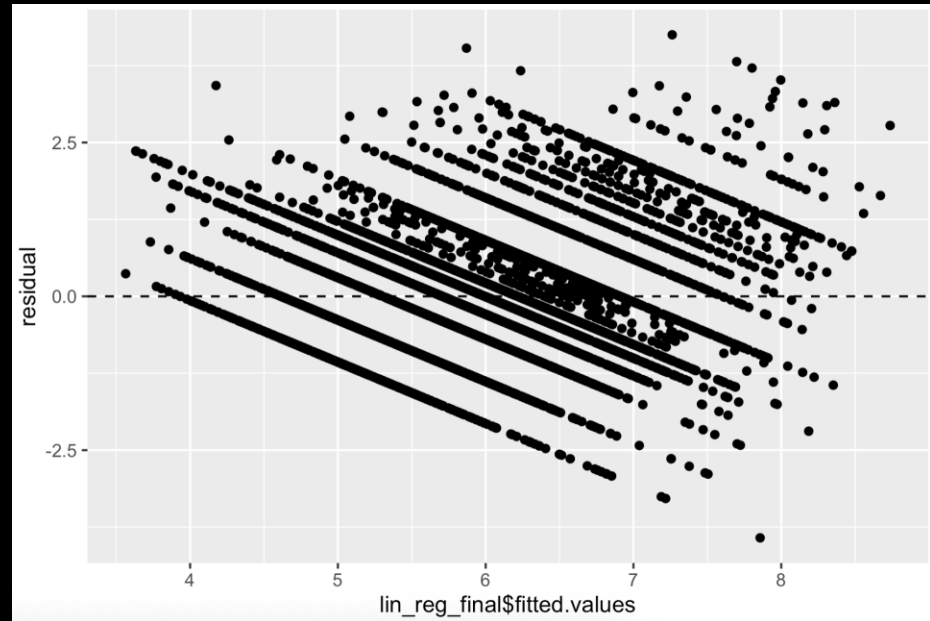


Linear Regression— Predicting Sales Volume

Our implementation of Linear Regression Model predicted products with high "Sales Volume" using units sold as the outcome,

- Gaming Products
- Sound Systems
- Laptops
- Batteries

Factors of high significance were, "Best Seller tag", "Rating, reviews", "Discounts" and "Sponsored Label"



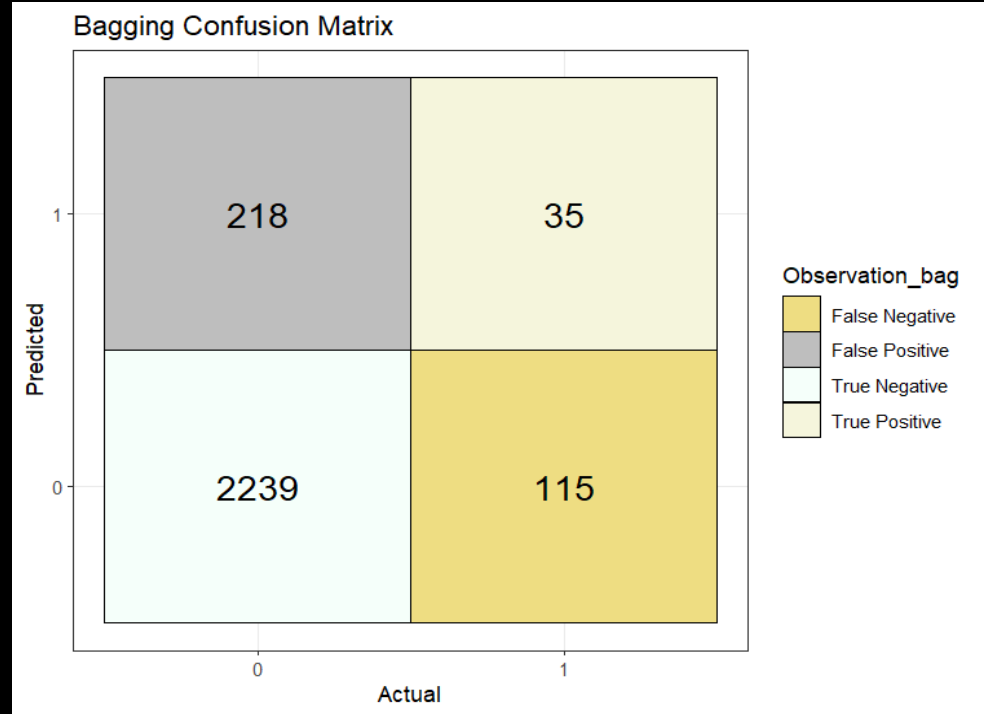
Logistic Regression - Predicting Best Sellers

Our implementation of Logistic Regression Model predicted whether a product will be a "**best seller**" using 6 predictors. The variables with high significance were,

- Price
- Sponsorship
- Units Sold
- Fast Moving Consumer Electronics

Confusion Matrix of Bagging Tree Model

- Used log values to counter Skewness
- High Specificity (91.1%)
- Reasonable Performance given imbalance
- Sensitivity is 23.3%
- Narrow Confidence interval (0.8588, 0.8848)



Model Evaluation

Model	Accuracy	Sensitivity	Key Notes
Logistic Regression	92.9%	17%	Highest Accuracy
Tree	94.4%	7%	Overfits
Bagging	87.2%	23.3%	Sensitivity Gain

Metrics Considered:

Accuracy: proportion of correct predictions.

Sensitivity (Recall): proportion of actual bestsellers correctly identified.

Sensitivity is the most important metric; we don't want to miss potential bestsellers!

Strategic Recommendations

- **Prioritize Positive Ratings:** Better ratings = higher sales.
 - 1 point increase in ratings ~2.7 higher odds of being Best Seller.
- **Sponsorship:** Sponsoring a product results in 4x higher odds of being Best Seller.
- **Optimize Discount:** For every percentage point discounted, odds of achieving Best Seller improve by 1%.
- **Build Reviews:** More reviews = Higher sales, increased odds of Best Seller.

Limitations

- **Survivorship Bias:** Dataset includes only currently listed products. Failed products that were delisted are not represented, potentially overestimating success factors.
- **Data Over Time:** Data represents a single time snapshot. Seasonal effects, economic conditions, and platform algorithm changes are not captured.
- **Causality vs. Correlation:** Our observational data cannot definitively establish causation.
- **External Validity:** Findings are specific to Amazon's electronics category. Generalization to other categories or platforms requires validation.

Model Limitations

- **Class Imbalance:** Best Sellers represented <6% of the dataset, resulting in model favoring the majority outcome.
- **Richer Features:** More features with strong correlations would paint clearer picture and improve predictions.
- **Data Imputation:** While this approach is supported, it may not capture the true values of the feature.
- **Reverse Causality:** Units sold and Best Seller status – which one is cause and which one is effect?

Suggestions for Future Analysis

- **Balance the Dataset:** Down-sample the majority class to regulate class distribution, or Up-sample the minority class.
- **Conduct A/B testing:** Determine if sponsorship, discounts, or other badges directly lead to Best Seller status.
- **Create AUC-ROC Curve:** Use graphical representation to find model that best predicts true positives and avoids false positives

Questions?

