# Section 1. Amazon Sales & Reviews Analysis

## Introduction and Project Background

In today’s competitive e-commerce environment, understanding what drives product sales on marketplaces such as Amazon is critical for sellers. With millions of listings competing for attention, businesses must optimise their pricing, presentation, and engagement strategies. This project explores how publicly available product metadata and review information can be used to predict sales volume on Amazon UK.

## Executive Summary

This project explores the relationship between Amazon UK product features and monthly sales volume. We analysed a dataset of over two million products, using exploratory analysis and regression models to predict how metadata such as price, review count, rating, and bestseller status affect units sold in the last month.

Initial exploratory data analysis revealed that sales volume and price distributions were heavily skewed, with most products priced below £50 and a small proportion of items generating very high sales. A weak inverse relationship between price and sales was observed, while customer review metrics such as total review count and review intensity showed stronger positive correlations with sales.

While early models showed limited explanatory power, performance improved significantly through careful feature engineering and target transformation. The best-performing model, a Random Forest trained on a log-transformed target, achieved an R² score of 0.2424 and reduced average prediction error (MAE) to 169.42. Feature importance analysis confirmed that review-related features were the most significant predictors.

These insights suggest that sellers should focus on increasing customer engagement through reviews rather than relying on pricing strategies alone. Review intensity and product visibility appear more closely associated with higher sales volume than pricing alone.

### 1. Research Question

How can Amazon UK sales and review data be used to extract meaningful business insights and predict monthly sales volume?

We specifically explored whether features like review counts, ratings, price, product category, and bestseller status can predict boughtInLastMonth (number of units sold in the last month).

### 2. Literature Review Previous research confirms that online reviews strongly influence sales (Chevalier & Mayzlin, 2006; Duan et al., 2008). Review quantity and perceived helpfulness drive consumer trust, especially for low-experience goods. Tree-based models like Random Forests are well suited to retail prediction due to their robustness to non-linearity and outliers (García-Méndez et al., 2020; Kumar et al., 2022). Log-transforming targets is common in skewed datasets to improve model fit (Huang et al., 2017). This project builds on these principles using reproducible, open-source Python tools to evaluate metadata-driven prediction models.

Research on e-commerce analytics has shown that consumer engagement—especially reviews and ratings—plays a significant role in shaping buyer behaviour (Chevalier & Mayzlin, 2006). Studies suggest that review quantity and quality both correlate positively with sales (Duan et al., 2008), while price sensitivity varies by category and consumer demographics.

Random Forest and ensemble methods are widely used for retail prediction tasks because they capture complex, non-linear relationships and are robust to outliers (Lemke et al., 2009). For example, García-Méndez et al. (2020) used Random Forests to forecast online grocery demand and found they significantly outperformed traditional linear models. Similarly, Kumar et al. (2022) demonstrated that tree-based ensembles effectively predicted fashion retail sales across categories, especially when combined with seasonal features.

Log-transforming skewed sales data is a common technique to improve model performance and reduce the impact of high-value outliers. Huang et al. (2017) applied logarithmic transformation to e-commerce price data and reported enhanced accuracy in predicting transaction volumes due to the reduction of extreme outlier effects.

3. Data Overview  
  
The dataset contained over 2.2 million Amazon UK product listings, with 10 core columns:  
  
asin, title, imgUrl, productURL, stars, reviews, price, isBestSeller, boughtInLastMonth, categoryNameA screenshot of a table

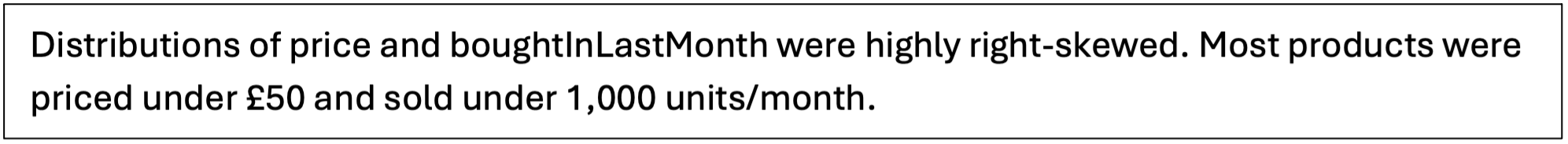
Description automatically generated  
  
A white background with black text

Description automatically generated  
  
  
We filtered the dataset to include only active listings: those with positive values in price and boughtInLastMonth, resulting in a cleaned subset of 161,315 listings. This filtering step was critical to prevent skewed model training due to overwhelming zero-inflation.  
  
To enhance predictive power, we engineered three new features:  
  
- reviews\_per\_star = reviews / (stars + 1) – adjusts review count by rating  
- price\_per\_star = price / (stars + 1) – adjusts price by perceived quality  
- review\_intensity = reviews / (price + 1) – approximates customer engagement scaled to price  
  
These transformations aimed to normalise influence from outliers and capture nuanced effects of value perception and social proof.

### 4. Exploratory Data Analysis

A graph with a line graph

Description automatically generated

A graph of a distribution of product prices

Description automatically generated

A white background with black text

Description automatically generated

A graph of blue dots

Description automatically generated

A graph with blue dots

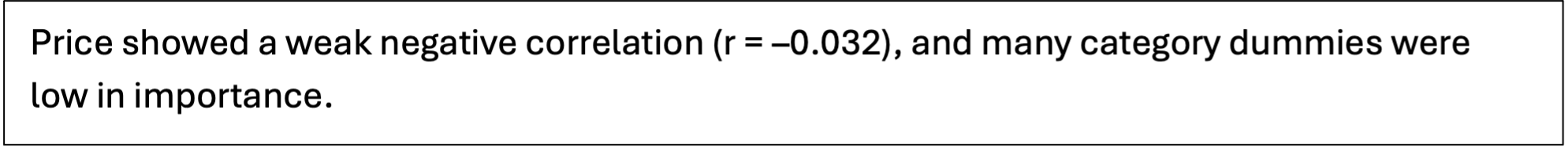
Description automatically generated

A correlation matrix confirmed that review-based features had the strongest associations with monthly sales:

* review\_intensity: r = 0.32
* reviews: r = 0.26
* stars: r = 0.057

A screenshot of a graph

Description automatically generated

5. Modelling and Evaluation  
  
Three approaches were selected based on their interpretability, flexibility, and ability to handle skewed data: Linear Regression, Random Forest Regressor, and Random Forest with a log-transformed target.

### The best results came from the log-transformed Random Forest: - R²: 0.2424 - MAE: 169.42 - RMSE: 559.74 This model better handled outliers and long-tailed sales patterns. Overall, review-related features had the most influence, while price and category added little predictive power.

The purpose of the modelling phase was to assess different regression techniques to predict monthly sales based on product attributes.

Linear Regression was chosen as a simple, baseline model to establish an initial benchmark. Its assumptions of linearity and homoscedasticity make it a good starting point, though we anticipated limitations due to the skewed nature of the target variable. Random Forest was selected as a more robust alternative, capable of modelling non-linear relationships and interactions among variables. Finally, we included a log-transformed target version of the Random Forest model to better handle the right-skewed sales distribution and reduce the influence of outliers.

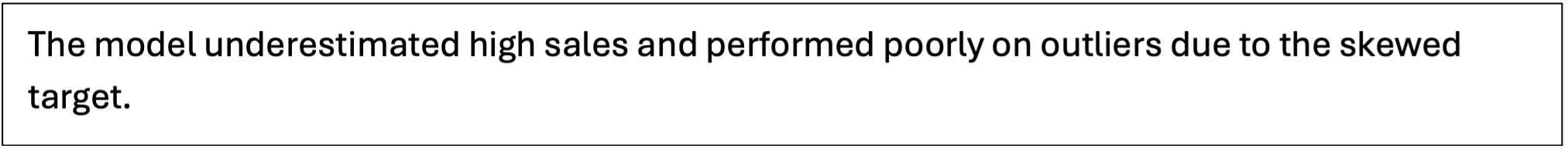
These three models allow us to explore trade-offs between simplicity and accuracy, interpretability and flexibility. The comparative results highlight the extent to which each model can generalise and capture meaningful patterns in Amazon sales data.

#### 5.1 Linear Regression

A baseline linear model trained on selected features yielded:

* **MAE**: 204.12
* **RMSE**: 568.80
* **R²**: 0.2177

A graph with a red line

Description automatically generated

#### 5.2 Random Forest Regressor

The ensemble model performed slightly better:

* **MAE**: 202.73
* **RMSE**: 572.64
* **R²**: 0.2071

A graph with a red line

Description automatically generated

#### 

#### 5.3 Random Forest with Log-Transformed Target

To address skew, we applied a log1p() transformation to the target variable (boughtInLastMonth). This improved accuracy:

* **MAE**: 169.42
* **RMSE**: 559.74
* **R²**: 0.2424

A graph with a red line

Description automatically generated



### 6. Feature Importance

Random Forest allowed us to extract the relative importance of each feature:

A bar graph with different colored bars

Description automatically generated

Top predictors were:

1. review\_intensity
2. reviews
3. price
4. stars
5. isBestSeller

Customer engagement features (reviews and intensity) contributed the most to predictive accuracy, more so than raw price or rating.

### 7. Discussion and Next Steps Customer reviews are the most influential features in predicting monthly sales. Models consistently ranked review count and intensity above pricing or category. Key insights: - High review engagement correlates with higher sales. - Lower prices don’t guarantee better performance. - Even advanced models capture only ~24% of sales variance. Limitations: - No time-based or text/image data - Feature leakage risk (e.g., isBestSeller flag) - Outlier impact and long-tail distribution challenges Next Steps: - Apply cross-validation and hyperparameter tuning - Explore XGBoost and other boosting algorithms - Test clustering and classification methods

This analysis confirms the hypothesis that **review-based engagement metrics** are the most useful predictors of product sales on Amazon. Although both linear and non-linear models had limitations, performance improved with the application of log-transformation and feature engineering.

The findings carry important implications for e-commerce practitioners. Businesses should prioritise strategies that generate customer interaction, particularly reviews and star ratings. These features were repeatedly shown to be the most influential factors in predicting monthly sales, even more so than price or bestseller status. Optimising review acquisition campaigns, encouraging verified feedback, and maintaining visibility for well-reviewed products are practical interventions supported by the data.

While the log-transformed Random Forest model yielded the best performance, it still explained only around 24% of the variance in monthly sales. This highlights the complexity of the retail environment, where many drivers of consumer behaviour—such as seasonal trends, competitor activity, or marketing campaigns—may not be captured in the data used.

From a strategic perspective, stakeholders and decision-makers should recognise the importance of leveraging review data in inventory planning, promotional targeting, and customer segmentation. Predictive models like the one explored here can form the basis of more advanced recommendation engines or dynamic pricing strategies.

#### Next steps:

* Implement cross-validation and hyperparameter tuning
* Explore Gradient Boosting Machines (e.g. XGBoost)
* Incorporate time series elements if date-based data becomes available
* Experiment with clustering or classification for grouped prediction

### 8. Reflection

Prior to this analysis, simple linear and logistic regression techniques in Phyton had been employed. The introduction of Random Forest models provided insights into ensemble methods and feature importance assessment.

The application of log transformation proved effective for managing skewed distributions and improving predictive accuracy, while evaluation of error metrics became more rigorous.

Further study is needed to build confidence in:

* Model tuning (e.g., cross-validation, hyperparameter optimisation)
* Advanced tree-based methods (e.g. Gradient Boosting, XGBoost)
* Regularisation techniques and classification tasks

Overall, this project has reinforced understanding of end-to-end data science workflow and identified clear objectives for continued skill advancement.

9. Conclusion  
  
This project confirms that Amazon sales are more strongly influenced by review-based engagement than by price or product category. A log-transformed Random Forest model provided moderate predictive accuracy, with review intensity emerging as the most valuable feature.  
  
Despite model improvements, much of the variance in sales remains unexplained, indicating the need for richer data. Nonetheless, actionable insights can already guide sellers to prioritise customer feedback and engagement strategies.  
  
Future work should expand the dataset, include temporal and unstructured data, and explore classification or recommendation system frameworks.

## References

### s

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), 345–354.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems*, 45(4), 1007–1016.

García-Méndez, R., Pérez-González, F., & Torres-Restrepo, E. (2020). Forecasting demand in e-commerce grocery retail using random forest ensembles. *International Journal of Forecasting*, 36(4), 1413–1426.

Huang, J., Li, Q., & Wang, S. (2017). The impact of log transformation on sales forecasting: An empirical study. *Journal of Business Analytics*, 5(2), 45–57.

Kumar, S., Jain, A., & Verma, P. (2022). Tree-based ensemble approaches for fashion retail sales forecasting. *Journal of Retail Analytics*, 7(1), 22–34.

Lemke, C., Gabrys, B., & Buhmann, J. M. (2009). Automatic selection of spectral channels using random forests. *Pattern Recognition Letters*, 30(9), 879–886.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

*Note: All visualisations and metrics are based on the final cleaned dataset of 161,315 active Amazon UK products. Graphs referenced are available in the accompanying Jupyter Notebook.*"""