Predicting Repeat Purchases in E-Commerce

Ayrton Stein
ayrtonstein513@gmail.com
Halifax, Nova Scotia, Canada
February 1 / 2023

ABSTRACT

As companies move their business online, the e-commerce space continues to develop as a rapidly growing industry. Marketing and upselling are crucial aspects of profitability. Therefore the objective of this investigation is to determine how the price, along with other features, of items sold on Amazon affect the likelihood of a customer making a repeat purchase. The underlying problem here is identifying more efficient upselling strategies for retailer companies. These findings may potentially help a retail company determine the likelihood of a customer purchasing another item based on their first purchase, and then decide how to upsell based on that information. Customer behavior and motivation is a crucial aspect of repeat purchases and original purchases in general, this notion has to be understood before attempting to answer the question. Identifying which features held the most weight in predicting repeat purchases was also a crucial aspect in determining if customers will buy again. In general it could be deduced that price, product brand, and product rating carry a great deal of weight when it comes to customer's making repeat purchases.

1.0 INTRODUCTION

The driving force behind a customer partaking in any kind of purchase varies from product to product. For example, habitual purchases such as toilet paper and toothpaste. In this case the consumer spends little time thinking or considering the purchase, rather they go with the brand they have always bought. Another example would be expensive purchases such as cars and luxury items, which are completed less frequently, and given more thought. However, when a customer is making a repeat purchase after having already bought something, the motivation for such purchase is completely different than that for the original purchase. This notion of consumer nature and motivation is the root problem that marketing agencies, and retailer companies work to overcome and capitalize upon. The purpose of this research is to determine how the price, along with other features, of items sold on Amazon affect the likelihood of a customer making a repeat purchase. This research problem is significant for a variety of reasons, and is applicable to multiple stakeholders in the space. From the retailer's perspective, this research and other investigations into the field have the potential to determine the likelihood of somebody becoming a repeat customer versus a one off, aiding retailers to construct more efficient upselling strategies. An examination into the nature of repeat purchases may also educate customers on how they are manipulated to buy additional products, and help them be more responsible consumers that are mindful of how companies create further business.

The notion of predicting repeat purchases deals in the realm of numerical data; supervised learning and classification models, specifically Random Forest, Neural Networks, and Decision Trees, are used to assess and answer such problem. A Logistic Regression model was also implemented however it logged the poorest performance. The resulting values after each model has been tuned and applied are numerical quantities. A model may predict a zero or one, for a completed repeat purchase or not, respectively. Accuracy percentages and f1 scores are used to assess the effectiveness of each given model in its ability to successfully predict if a given customer did or did not purchase an additional item.

2.0 BACKGROUND - LITERATURE REVIEW

A literature review was conducted in order to achieve a greater understanding of the problem at hand, and possible solutions to similar problems that have already been proved successful. A study conducted by researchers in India sought to examine a similar problem to that of this research paper. Such study sought to predict repeat customers on e-commerce sites based on both the number of purchases in the last year (aggregate), and the time series of daily-weekly purchases (Temporal), Agarwal, Shroff et al used Long short term memory for the temporal aspect and Quantile regression for the aggregate aspect, as their classifiers (Agarwal). They found that the

temporal version of data picked up on more aspects of customer behavior versus the aggregate and quantile regression model, in general the long short term memory model more accurately distinguished between non-repeaters and repeaters (Agarwal). Most existing models that seek to predict repeat customers only use data relating to purchases over the last year, and not the time series, therefore this experiment was unique in including an additional model in the LSTM, and Agarwal, Shroff et al even found this additional model to be more accurate (Agarwal). Another study conducted by Sifa, Hadiji et al sought to research this field under the umbrella of mobile games. The study essentially found that the driving factors for purchases were play time, country, and how far players manage to get in the game, both the general regression and classification models found similar results (Sifa). Sifa, Hadiji et al used a variety of models within the realm of classification and regression, of the ones used, Random Forest, Smote-NC, and Decision Tree, scored the highest accuracy which is insightful for predicting repeat purchases as well (Sifa). The information from these cases was insightful for this research, as both Random Forest and Decision Trees were used, because of their success in the similar cases outlined above.

3.0 DATASET INFORMATION - PREPROCESSING

"Amazon_Data", being the dataset used for this research experiment, was found on the kaggle database. It includes 14 features (See Table 1) both numerical and language based, and up to 10,000 samples. From the 14 features, seven were chosen to be included as indicators for the AI model's predictions. These seven features are as follows: price, number available in stock, new/used, average review rating, number of answered questions, number of reviews, and brand. These seven were chosen as they are the most significant in creating a positive or negative customer experience. For example a product purchased with a low average review rating may leave the buyer disappointed and unlikely to complete a repeat purchase. Additionally certain outliers existed of these features within certain samples. Data points with prices exceeding 1000 dollars were excluded, as well as certain points that contained null values for multiple features. These outliers distorted scatter plots and histograms, and were therefore removed from the dataset. In order to understand how these features related and correlated with one another scatter plots and histograms were made to plot certain features against one another (See Figures 1-2 for examples). Of the seven features listed, some of them included both numbers and letters in their values; because of how the AI model's interpret these values, numbers are the only suitable option, therefore letters were removed from these specific seven features' values. For example, the currency symbol had to be removed from the price values, and each distinct brand of product had to be assigned a number value. Additional columns were created to include these numerical values with the word new preceding its original name. As for implementing the dataset into the AI models, an 80-20 train-test split was used in order to maximize the accuracy.

Features Included in Study	**Features Excluded from Study**
New average review rating	Uniq id
New/used	Product name
New classifier brand	Manufacturer
New price	Price
New number available in stock	Number available in stock
Number of reviews	Customer questions and answers
Number of answered questions	Product description
	Product information
	Description
	Amazon category and sub category
	Average review rating
	Items customers buy after viewing this item

Table 1. Index of Features Included in Amazon_Data

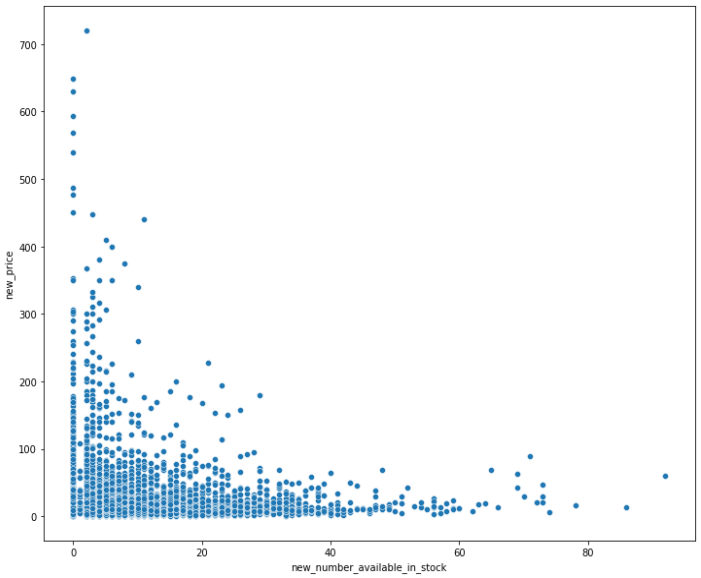

Figure 1. Scatterplot: Number Available in Stock versus Price

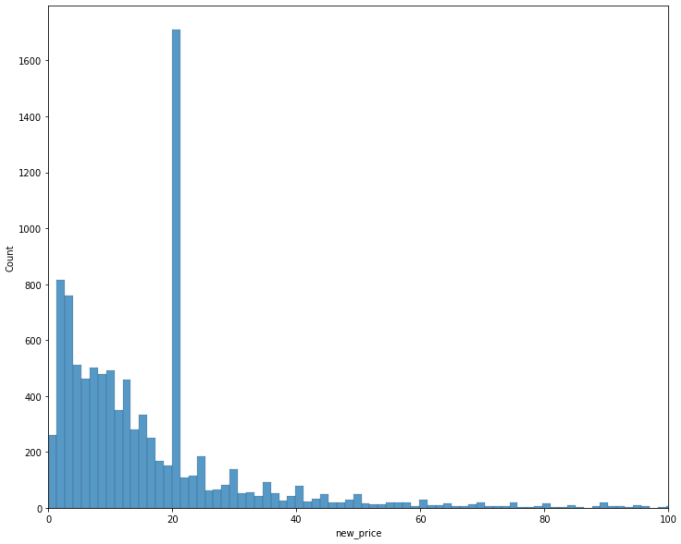

Figure 2. Histogram Plot: Price versus Count

4.0 METHODOLOGY - MODELS EMPLOYED

In order to go about determining how the price, along with other features, of items sold on Amazon affect the likelihood of a customer making a repeat purchase, a background of understanding was necessary, the process for which is outlined in the previous section of 2.0. Processing and cleaning of "Amazon_Data" was another necessary step which additionally is outlined in section 3.0. Amazon_Data was split into a train/test split of 80/20 respectively. After these first two steps were completed, a Logistic Regression model was applied to the dataset, as a baseline of success for the subsequent models that would be employed. Logistic Regression is a linear method for binary classification problems, which uses the logistic function to compress the output of the linear combination of the inputs into a probability between 0 and 1. This probability can then be thresholded to obtain the final binary class prediction.

A Random Forest Classifier was then applied to the training data. A Random Forest Classifier is an ensemble machine learning algorithm that combines multiple decision trees to predict the class of a given data point. Following the use of the Random Forest Classifier, a Decision Trees model was utilized. A Decision Tree is a tree-structured machine learning algorithm that is used for classification problems. It works by recursively splitting the data into subsets based on the values of the features, creating a tree-like model of decisions and their possible consequences. Finally a Neural Network model was applied to the training data. A Neural Network is a machine learning model inspired by the structure and function of the human brain. It consists of layers of interconnected nodes, called neurons, which process and transmit information.

In the pursuit of maximizing each models' accuracy, an element of hyperparameter tuning was also included. A code which tested a range of values for the given hyperparameters of each model was developed and used to identify the highest possible accuracy. A depth of 1 versus 10 nodes, for example, in the Decision Trees model would affect the resultant accuracy, and therefore a maximizing value had to be found. This process of hyperparameter tuning was done on the Decision Trees, Random Forest, and Neural Network models. The Logistic Regression was not subjected to hyperparameter tuning as it was merely used as a baseline of success. After each of these four distinct models were applied and tuned, their respective accuracies were scaled to a 1-100 score, and confusion matrices were generated to visualize each models' strengths and weaknesses.

5.0 RESULTS AND DISCUSSION

The Logistic Regression model was included as a baseline of success. It was therefore not expected to have the greatest resulting accuracy or metrics. Despite this, the Logistic Regression model performed just as well as the other models, with an accuracy percentage of 70.91%, this percentage refers to the rate of accurate predictions of repeat purchases by the respective Ai model. The Logistic Regression model also had an F1 score of 0.8282. This F1 score is a statistical measure of the model's accuracy that balances precision and recall. It takes into account both the false positives and false negatives of a model and calculates the harmonic mean between precision and recall. The F1 score ranges from 0 to 1, with a score of 1 indicating perfect precision and recall, and a score of 0 indicating the worst possible performance. Secondly, the Random Forest model which performed the best of the four, had an accuracy score of 72.11% and an F1 score of 0.7967. The Decision Trees model performed worse than the Logistic Regression model, with an accuracy of 70.81% and an F1 score of 0.8237. Finally the Neural Network model performed the most poorly, as it had an accuracy of 70.76% and an F1 score of 0.8119. In order to achieve these results, and to maximize the accuracy, the hyperparameters of each model were adjusted and tuned. Table 2 showcases each models' metrics and the hyperparameters that achieved such values.

Model Name	Accuracy (%)	F1 Score	Hyperparameters
Logistic Regression	70.91	0.8282	-

Random Forest	72.11	0.7967	Max Depth = 11 # of Estimators = 90
Decision Trees	70.81	0.8237	Max Depth = 7
Neural Network	70.76	0.8119	Hidden Layer Size = 4 Max Iterations = 200

Table 2. Each Models' Metrics and Associated Hyperparameters

An important aspect of each models' accuracy is the ratio between false positives/negatives and true positives/negatives. This information is shown in confusion matrices, which provides a summary of the model's accuracy by comparing the predicted class labels to the actual class labels. The confusion matrix displays the number of true positive, false positive, false negative, and true negative predictions made by the model. Figures 3-5 demonstrate the given confusion Matrices for the Random Forest, Decision Trees and Neural Network models respectively. A confusion matrix was not included for the Logistic Regression Model, as it was meant to simply act as a baseline of success.

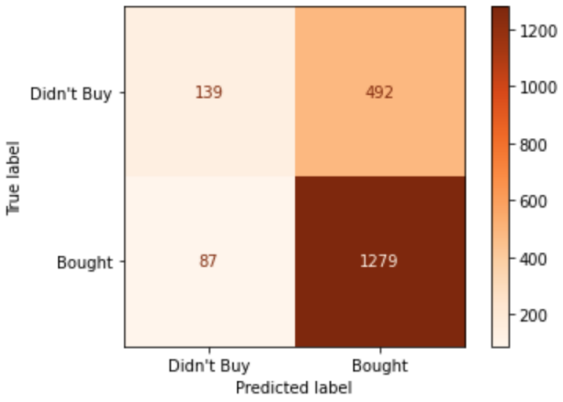

Figure 3. Random Forest Model Confusion Matrix

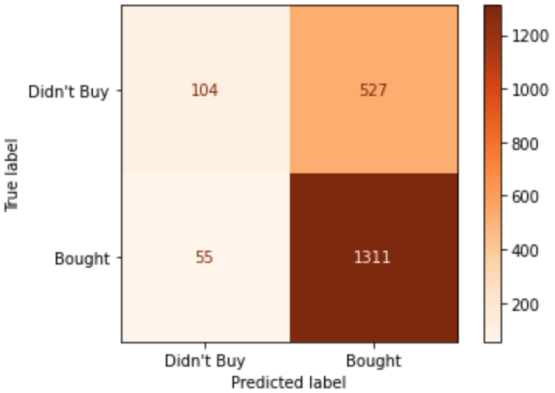

Figure 4. Decision Trees Model Confusion Matrix

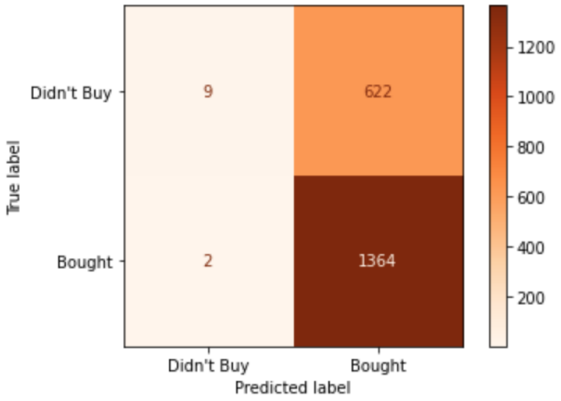

Figure 5. Neural Network Model Confusion Matrix

The accuracy of each model is greater than 50% meaning the model is performing better than random guesses. This proves that there is a true correlation between the chosen features (See Table 1) and the likelihood of repeat purchases. However Figures 3-5 demonstrate the ratio between the positive and negative predictions each model is making, and in all cases the majority of the models' errors are in predicting "Bought" when the true value is "Didn't Buy". This bias in guessing is resulting in a relatively high accuracy score, because of the fact that there are more repeat purchases than not within Amazon_Data itself. This is a potential flaw in the models, as an aspect of each models' predictions stems from the notion of the true value being a repeat purchase every time. This is especially true for the Neural Network model, which is understandable as it also performed with the poorest accuracy of the four. This phenomenon is less prevalent in the Random Forest model's confusion matrix, which is reassuring as this model performed the best.

Figure 6 shows the Receiver Operating Characteristic (ROC) Curve for the Random Forest Classifier model. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The ROC curve helps to visualize the trade-off between TPR and FPR, as the threshold setting is adjusted. A perfect classifier would have an ROC curve that passes through the upper left corner meaning a 100% TPR. This curve is included for the Random Forest model as it provides valuable information about the model's performance.

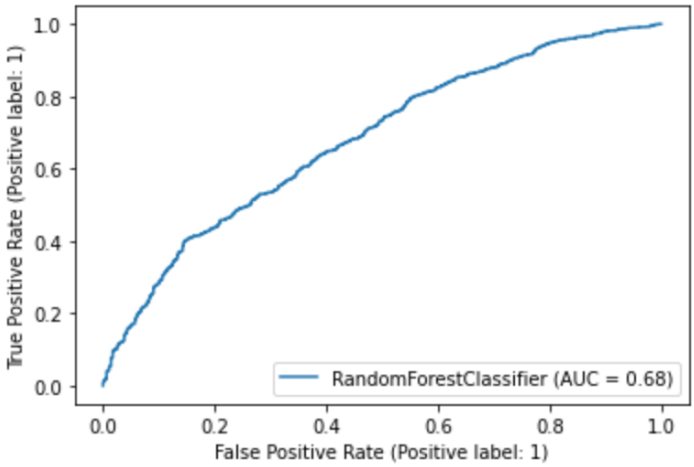

Figure 6. ROC Curve: Random Forest Model

The objective of this research is to determine how the price, along with other features, of items sold on Amazon affect the likelihood of a customer making a repeat purchase. Therefore an understanding of the weight each feature holds with regard to the resulting prediction of the given AI model is a necessary aspect to the conclusion and resolution of this research. Table 3 showcases the respective weight of each included feature, with a value between 1 and 0: A higher value meaning a greater level of influence.

Feature	Feature Importance
New Price	0.3031
New Classifier Brand	0.2831
New Average Review Rating	0.1564
Number of Answered Questions	0.0811
New Number Available in Stock	0.0774
Number of Reviews	0.0765
New/Used	0.0225

Table 3. Features and their Respective Importance

As demonstrated in Table 3, the average review rating, brand, and price, in increasing order, are the most influential features in predicting the likelihood of a customer completing a repeat purchase. These results directly answer the Research Question, that being: How the price, along with other features, of items sold on Amazon affect the likelihood of a customer making a repeat purchase. In conclusion, the average review rating of a product, its brand, and its price affect the likelihood of a customer making a repeat purchase to the greatest extent.

6.0 CONCLUSION

The purpose of this research is to determine how the price, along with other features, of items sold on Amazon affect the likelihood of a customer making a repeat purchase. This question and others of similar nature are significant because they create the potential ability to determine the likelihood of somebody becoming a repeat customer versus a one off: aiding retailers to construct more efficient upselling strategies. An examination into the nature of repeat purchases may also educate customers on how they are manipulated to buy additional products, and help them to be more responsible consumers.

It was concluded that average review rating, brand, and price of a given item hold the most leverage in creating a repeat customer or one-off buyer. These findings provide a direct answer to the question. All four of the incorporated models performed with an accuracy significantly greater than that of a random guess which proves there is a real correlation between the average review rating, brand, and price of a given item, along with the other included features (See Table 1), and the likelihood of a repeat purchase. Applying the Random Forest model to different datasets, as it was the most successful of the four, may help to solidify the correlation that has been identified here. Additionally, further hyperparameter tuning on a more advanced computing system would improve the accuracy of the models' predictions and perhaps promote further consistency which would manifest in a confusion matrix.

Acknowledgments

Thank you to Bryce Johnson for your guidance and mentorship throughout this process.

References

[1] Agarwal, Auon Haidar Kazmi Puneet, et al. *Deep Temporal Features to Predict Repeat Buyers*. TCS Research, www.researchgate.net/profile/Auon-Kazmi/publication/292972291_Deep_Temporal_Features_to_Predict_Repeat_Buyers.pdf. Accessed February 4, 2023

[2] Patel, Neil. "How To Upsell Any Customer." *Forbes*, 21 Dec. 2015, www.forbes.com/sites/neilpatel/2015/12/21/how-to-upsell-any-customer/?sh=75f0cca1c406.

[3] Sifa, Rafet, et al. *Predicting Purchase Decisions in Mobile Free-to-Play Games*. Fraunhofer IAIS, https://ojs.aaai.org/index.php/AIIDE/article/view/12788/12636. Accessed February 4, 2023

[4] "Why Marketers Need to Pay Attention to Consumer Behavior." *Skai*, Skai TM, 9 May 2022, https://skai.io/blog/consumer-behavor/?utm_source=google&utm_medium=cpc&utm_campaign=Skai_2022_DSA_Blog_EN