Machine Learning Model to Predict the Best Seller Rank of new Amazon Listings

Thomas Robinson supervised by Oliver Vogt and Stefano Giani

Abstract—**The success of a product on Amazon is heavily influenced by the contents of its listing. Along with a product's price and rating, it is the semantic information portrayed in its listing images and description which are most important. However, previous studies in this area have neglected semantic information as it is difficult to quantify. This report details the use of state of the art machine learning models, which enable semantic interpretation, to build a model predicting the 'best seller rank' (BSR) of product listings on Amazon. The model achieves an accuracy of 57% for predicting 6 categories of BSR. The report also provides an improved regression model predicting product sales from BSR with an R^2 of 0.75, establishing the relationship between BSR and sales. Finally, the implementation of this model as a tool to be used by sellers is demonstrated, and suggestions for further improvements are provided.**

I. Introduction

AMAZON is the largest and fastest growing ecommerce company in the western market, currently accounting for "39.5% of all US retail ecommerce sales in 2022", with that Figure growing by approximately 1% each year [1]. As well as this "66% of consumers start their product searches on Amazon" [2], showing that Amazon has become their default online shopping place. It is therefore not surprising that "roughly 3,700 new sellers join Amazon every day" [2].

However, it is not easy for new sellers to take advantage of Amazon's market dominance. "According to Amazon: 70% of Amazon consumers rarely navigate through the first search results page" and "35% tap on the first item listed on the search page" [3]. The order in which a list of products relating to a search appears is generated by the Amazon A10 search algorithm. This algorithm promotes products that have generated more sales to be seen first, as these are perceived as the best products [4]. It also ranks products in their given department and sub-departments. This ranking is known as the 'best seller rank' (BSR). Sales, specifically recent sales, are the main factor affecting a product's BSR, therefore the BSR is a good metric for the performance of a given product [5]. While this algorithm is effective at promoting what are perceived to be the best products, it does heavily favour a small number of products which are already doing well, meaning newer sellers with new products may struggle to get noticed in the market.

Consumer's purchasing decisions depend on their perception of the quality and value of the product [6]. This perception can be influenced by a multitude of factors, including positive customer reviews, the BSR itself, and opinions of the brand and product [7]. The page which a customer views to learn about and therefore forms an opinion of the product is known as it's 'listing'. Therefore features of the listing can be used to predict the sales rate and BSR of a product [8].

How a listing influences a customer's opinion is however very complex [8]. It depends on both conscious and subconscious interpretations of the physical information being conveyed and the semantics behind it. New machine learning models are very effective at interpreting complex relationships, and novel models can even extract semantic data from images and text, and so could be used to aid in the prediction of product listing performance [8, 9]. As well as this they can be used to provide feedback to the seller to inform them of ways to improve their listing [10].

This report will outline previous attempts to predict product performance on Amazon, and the creation of a model which builds on these. The model will use state of the art machine learning methods to better predict the BSR of Amazon products from the contents of their listing. It will also provide feedback indicating improvements that can be made to improve this BSR.

II. Review of related literature

This section will critically analyse strengths and weaknesses of previous studies using machine learning to predict the performance of Amazon listings. It will provide: an overview of the previous related studies; an analysis of the listing features used in these studies and novel methods for extracting more useful features; an analysis of the methodologies of these studies and how they could be improved; and finally, a summary of the biggest improvements that can be made and how this model will address those.

A. Overview

Some studies have attempted models analysing listings to predict performance. [10] used an ensemble algorithm for categorising listings into "best-selling", "non best-selling", and "worst-selling" products. Description text was "vectorized" using a "term-frequency inverse document-frequency" approach, which analyses individual words seen in text without context. Images were analysed using a Convolutional Neural Network, (CNN) and categorised as a binary positive or negative. The best ensemble model achieved an accuracy of 83.3%, however in the largest department, "Home and Kitchen", the random forest regressor was only outperformed by 0.6%. [10] states that more advanced text and image analysis would likely lead to improved results.

Another study, [8], used even more data in their model, for predicting "conversion rate" of Amazon landing pages. Their analysis of individual areas like text and images was more limited, the only features used being description length and number of images. However, by including features such as historic conversion rate, marketing budget, and number and average rating of reviews they were able to reduce their average error to 27%. The model architecture used was gradient boosted decision trees. This demonstrates the effectiveness of using a wider range of features, which allow for more information to be included in the model. They also provided users with a tool providing suggestions for how to improve their listing in each area of their model, with feedback indicating an average satisfaction of 5.8 out of 7 [8].

Further academic research on the creation of tools to improve a listing's best seller rank is lacking, and current commercially available tools are basic. *Jungle Scout* offers a tool which analyses title length, number and length of features (bullet points), description length and number and resolution of images [11]. It does also, however, incorporate a keyword and back-end search term analysis, which specifically optimises for Amazon's A10 search algorithm [11]. Another tool, provided by SellerApp, offers a similar detail of analysis, yet claims a 67% conversion rate improvement, and an 83% increase in organic traffic. This highlights the potential efficacy of a more sophisticated tool [4].

Another study focused more on the direct link between BSR and sales [5]. Using a regression algorithm for predicting the logarithm of the sum of sales, over a 3 month period, using the BSR, they were able to achieve an R^2 value of 69%. Indicating their model accounted for 69% of the variability in the sales. They state that using a more complex model with more features could increase the accuracy [5].

B. Features

This section will provide an in depth study into previous uses of each of the features of an Amazon product listing for predicting BSR. It will also detail the potential that state of the art machine learning techniques have for this.

1) Description: The product description is a key area of importance on a listing, where the consumer will get most of the item's technical information to make purchasing decisions [4]. [8] simply used the length of the description in their model to provide a proxy for the amount of information it contained, and still found this to be useful. However, it is the detail of the words and sentiment in the description that truly influential [12]. [10] attempts to analyse this sentiment, however their model treats each word as it's own feature, therefore has no interpretation of context, and is simply looking for good key words to use. A similar study looking to analysing descriptions on ecommerce site *Rakuten*, used a Recurrent Neural Network (RNN) with a Long Short-Term Memory (LSTM) architecture to interpret 'embedded narratives' in the text. They found that including the text data with embedded analysis in their models led to an improvement in R^2 of up to 0.34 [12]. The latest Natural Language Processing (NLP) models, for example *OpenAi's GPT-4*, use a transformer architecture for even greater accuracy, and could be used to further improve model accuracy [13].

2) Images: Images are another key aspect of the listing, where the consumer can get a feel for the quality of the product visually [4, 14]. The number of images was the only image metric used by [8] but was still found to be useful. In other studies, increasing the number of images included has been shown to be "an effective way to improve sell-through", by providing a "more complete visual representation of the product" [15]. Extracting visual features such "brightness", "contrast" and average "lightness" can indicate the quality of images, inferring the quality of the product to the customer. [16] found including such image features in their model for predicting clicks in an "ad auction" improved their area under curve (AUC) metric by 6.5%; and [17] found including theirs in prediction of *Etsy* listing "popularity" improved their AUC accuracy by 3.45%. However, these features don't directly contain any semantic information from the image that a human would see.

More novel machine learning techniques have been used to extract semantic information from images. The model by [10] used a CNN to categorise the images as positive or negative. They however only found this method to be effective for listings of books, suggesting that the model struggled to interpret useful information when images of a category were significantly different [10]. [9] used open source pre-trained CNN models to assess images in an "online catalog" and optimise which images are shown and in which order for a product. Their optimisation was found to have a "Relative effect on Add-to-Cart" of 27% in the monitors category, showing how effectively more advanced machine learning can interpret images. More recent advances in image machine learning also offer opportunities. *OpenAi's* popular *DALL.E* image generation tool is powered by Contrastive Language–Image Pre-training (CLIP). This provides and compares embeddings of semantics from images and text extremely effectively, and could be used to extract semantic data from landing page images [24].

3) Reviews: Reviews have been shown to be a key predictor in the success of Amazon listings, and ecommerce performance in general. They offer "social information", which [18] found to be a "driving factor behind online consumer choice". Reviews on Amazon contain a rating out of 5, and in some cases a worded review. The model by [8] only took number of reviews and average rating into account, and the model by [10] excluded them. Other studies have found that "high ratings that are validated by a large number of reviews have a positive effect on sales" [19, 20]. However, the direction of causation between sales and review volume can be difficult to decipher – more reviews may simply indicate a more sales have already been made. The importance of "review sentiment" over simple ratings is debated. While the model of [21] "did not project these factors [sentiments] as important", [22] found "the impact of numerical ratings on sales being mostly indirect and through sentiments". Like with the product description, more advanced models could be used to analyse this sentiment in detail.

C. Methodology

This section will critically analyse the methodologies used in the previous related studies.

Useful data relating to Amazon sales is closely guarded. The BSR prediction model created by [10] used the publicly available 'SNAP Amazon Review Dataset', which contains data from the content of Amazon listings, such as the title, price, description, images and BSR. However, this data set only includes listings up to 2014, and is therefore outdated [10]. It also lacks time stamps, meaning changes in each department over time cannot be accounted for. This is especially relevant since the BSR can change significantly over short periods of time due to recent sales [10, 5]. The data used in the conversion rate prediction model by [8] addressed these problems, being collected up to 2020 and containing tracking over time. It also contained transaction data, meaning the model could relate to product sales.

While the data used by [8] was of higher quality than [10], their analysis of the listing itself was much less detailed, as mentioned in the features analysis section. The data used by [8] also didn't allow a link between BSR and product sales to be found. Therefore, a study relating detailed listing features both BSR and sales is lacking.

D. Summary and Focus of Work

Analysis of currently available literature relating product listings to their performance has been outlined. The BSR has been predicted from product listing features with some success, however this has been done with both outdated data and machine learning techniques. Some studies have used their models for the creation of tools to be used by sellers, which have proven to be popular. More advanced techniques able to interpret sentiment from images and text have been shown to be effective in other areas, such as newer NLP with LSTM and CLIP image embedding and could be applied to a listing model. While a simple model relating BSR to sales was created, it was found that using more features and information may enhance this model.

The focus of this work will therefore be to utilise modern machine learning techniques to analyse the sentiment of product listings in more detail, to better predict a their BSRs. This will need to be performed with more recently collected data in order to maximise its relevance. It will also provide feedback to the user, detailing strengths and weaknesses of a given listing. This has been demonstrated to be a useful and popular tool for new sellers, and the greater detail of this model will allow for more detailed and specific information to be conveyed. Finally, a model relating BSR to product sales will also be created, which will take into account more features from the listing than previous work, leading to greater accuracy.

III. Data

A. Listings Data Set

A data set relating listing features to BSR has been provided by *Apollo*. It contains parsed HTML content from approximately 35k listings across all Amazon departments, with 16k from the largest department - 'home & kitchen'. This data has been stored in the non-relational database *MongoDB*. Features of the listing available include the Amazon Standard Identification Number (ASIN), product name, copy-writing features (the bullet points and product description), the title, image and video data, price information, ratings information and the BSR in both the main Amazon department and its sub-department (for example 'home & kitchen' and then 'kitchen & bath fixtures').

The data was collected between August and December 2022, so is relevant to the current Amazon market. It includes all useful details from the listing, except the content of any reviews made. It was initially randomly collected from each department. However, due to the very low proportion of listings that do well (only about 10% even have a BSR), it was later targeted to better performing listings with lower BSRs. Still only 15% of the listings in the data set have a BSR, which means quantity of data may be a limiting factor for the model's accuracy.

The data was also stored in raw HTML form and contained some inconsistencies due to differences in Amazon's web pages. Therefore, some processing was needed to extract and correctly label useful features.

The distribution of department BSR is shown in Figure 1. It is generally uniform, with an increased representation of lower ranks, which is expected given the data collection methods. This will allow high performing listings to be represented as strongly as mid to low performing listings and should allow a balanced model to be produced. There are also 14k listings with no BSR.

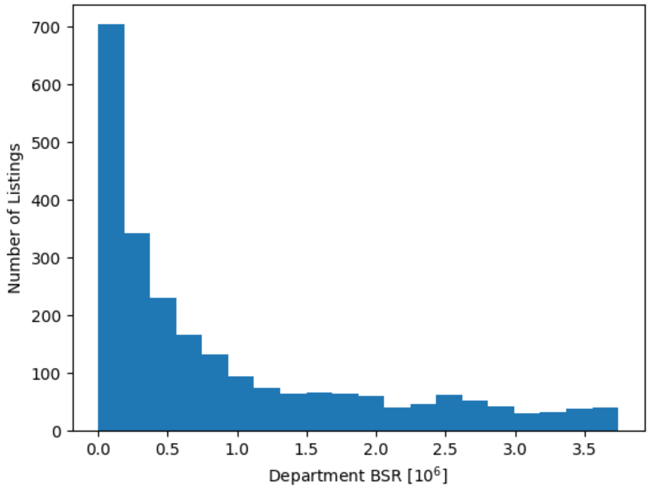
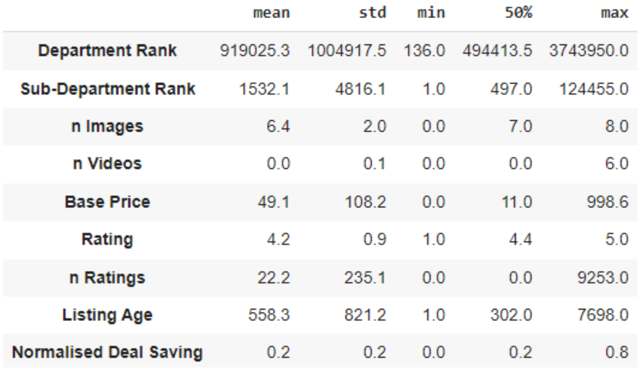


Fig. 1. Distribution of the department BSR in the listings data set.

A summary of the numerical data is shown in Figure 2. The department rank and sub-department rank are the BSRs in the department and sub-department respectively. The Normalised deal saving is any saving that has been offered, as a fraction of the total price of the product. It is seen that most listings have close to the maximum number of images which is 8, with a mean of 6.4 and a median of 7. It is also shown that very few listings have videos, suggesting this may not be a useful metric.

Figure 3 shows the distribution of the number of ratings for each listing. It shows how most listings have fewer than



	mean	std	min	50%	max
Department Rank	919025.3	1004917.5	136.0	494413.5	3743950.0
Sub-Department Rank	1532.1	4816.1	1.0	497.0	124455.0
n Images	6.4	2.0	0.0	7.0	8.0
n Videos	0.0	0.1	0.0	0.0	6.0
Base Price	49.1	108.2	0.0	11.0	998.6
Rating	4.2	0.9	1.0	4.4	5.0
n Ratings	22.2	235.1	0.0	0.0	9253.0
Listing Age	558.3	821.2	1.0	302.0	7698.0
Normalised Deal Saving	0.2	0.2	0.0	0.2	0.8

Fig. 2. Summary of numerical data in the listings data set

600 ratings, and a small number have many more. The mean number of ratings is just 22, as seen in Figure 2. This emphasises the effect that Amazon's algorithm can have on the market, where listings with a better BSR will be seen more and therefore purchased exponentially more than a listing with a poorer BSR.

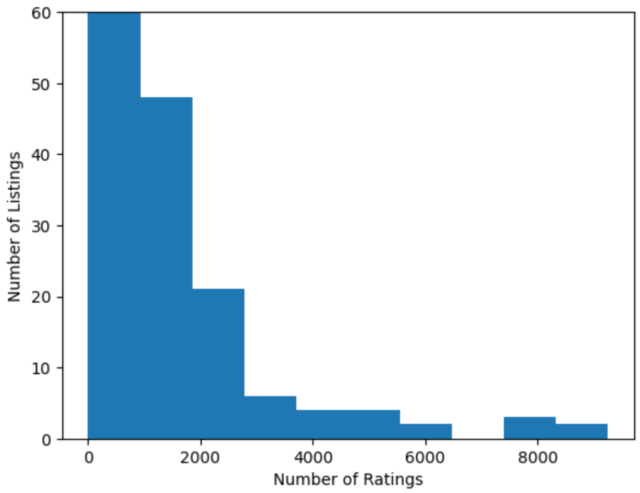
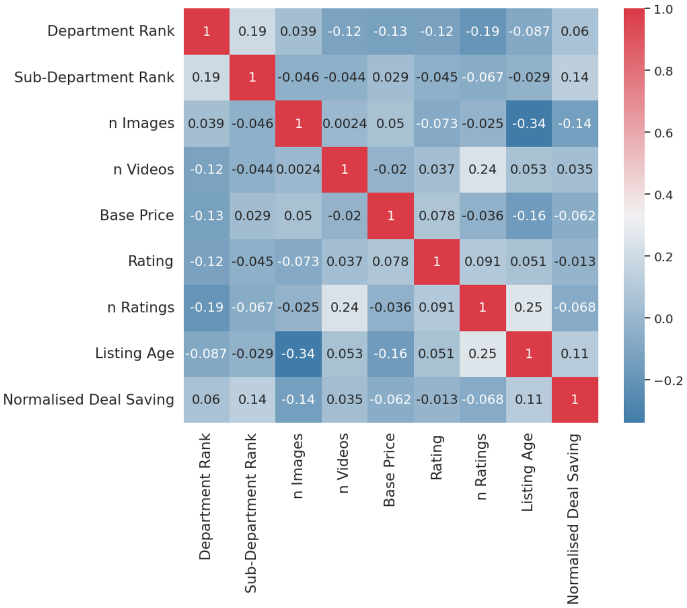


Fig. 3. Distribution of the number of ratings for listings in the data set. The y axis has been shortened for clarity - 16000 listings had between 0 and 600 ratings.

Figure 4 shows the correlation between different features of the listing. Negative correlations with department rank (department BSR) indicate performance improves as the feature increases, as the department BSR will decrease. The department rank shows a weak negative correlation with the number of videos, the price, the rating and listing age. There is also a stronger negative correlation with the number of ratings. This is expected as it is an indication of a large volume of sales, therefore these variables are not independent. Interestingly both number of images and deal saving have shown weak positive correlations with department rank, which is the opposite of what would be expected. For images, this may be due to most listings being close to filling the maximum of 8 images, suggesting that the quality of the images may be more important than the quantity. The positive deal saving correlation may be because when a product does not perform well sellers may then offer a deal to increase sales, so these listings are often further from the top.



	Department Rank	Sub-Department Rank	n Images	n Videos	Base Price	Rating	n Ratings	Listing Age	Normalised Deal Saving
Department Rank	1	0.19	0.039	-0.12	-0.13	-0.12	-0.19	-0.087	0.06
Sub-Department Rank	0.19	1	-0.046	-0.044	0.029	-0.045	-0.067	-0.029	0.14
n Images	0.039	-0.046	1	0.0024	0.05	-0.073	-0.025	-0.34	-0.14
n Videos	-0.12	-0.044	0.0024	1	-0.02	0.037	0.24	0.053	0.035
Base Price	-0.13	0.029	0.05	-0.02	1	0.078	-0.036	-0.16	-0.062
Rating	-0.12	-0.045	-0.073	0.037	0.078	1	0.091	0.051	-0.013
n Ratings	-0.19	-0.067	-0.025	0.24	-0.036	0.091	1	0.25	-0.068
Listing Age	-0.087	-0.029	-0.34	0.053	-0.16	0.051	0.25	1	0.11
Normalised Deal Saving	0.06	0.14	-0.14	0.035	-0.062	-0.013	-0.068	0.11	1

Fig. 4. Correlations between numerical metrics in the data set.

B. *Sales Data Set*

Sales data has also been provided by *Apollo*. This was in the form of 19 *Facebook Business Reports*. Data in these reports included the product ASIN, name, units ordered, and sales made in dollars among other ecommerce business metrics. This data was a total over the previous month, which was October 2022. A data set was also provided which included all listing details for each product, in the same format as the previous listing data set.

A large proportion of this data set were different items from the same listing (for example the same t-shirt in different sizes). The average and sum of this data from each independent listing was aggregated, then combined with the listings data. After aggregation, data from 217 listings remained, of which 122 had no fields missing. The sales data's correlations with listing features were investigated and are shown in Figure 5. The mean of product sales was chosen as the most appropriate metric, as it had the most business relevance, has been used in previous studies and had the highest correlation with BSR.

IV. Methodology

This section outlines the creation of each model and processing the data available into useful features.

A. *Data Processing*

The raw data had to be processed before being used by the model. This included extracting features from the images, evaluating the effectiveness of the product description, accounting for unranked listings and labelling categorical data such as sub-department and ratings.

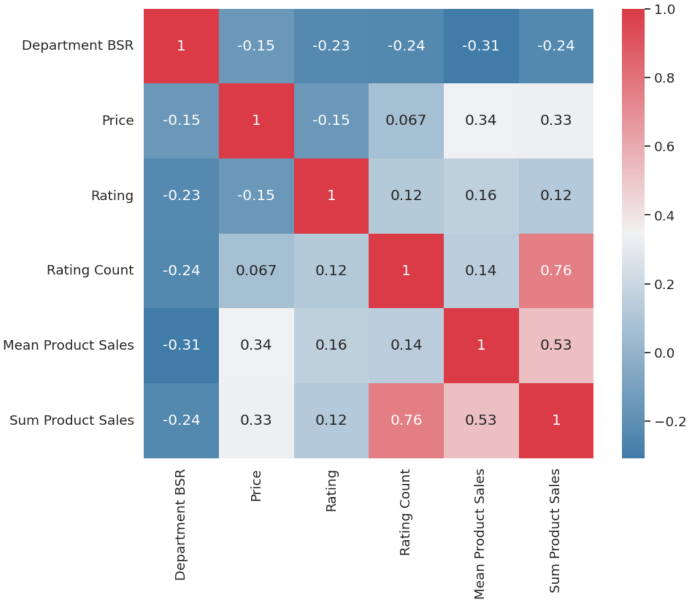
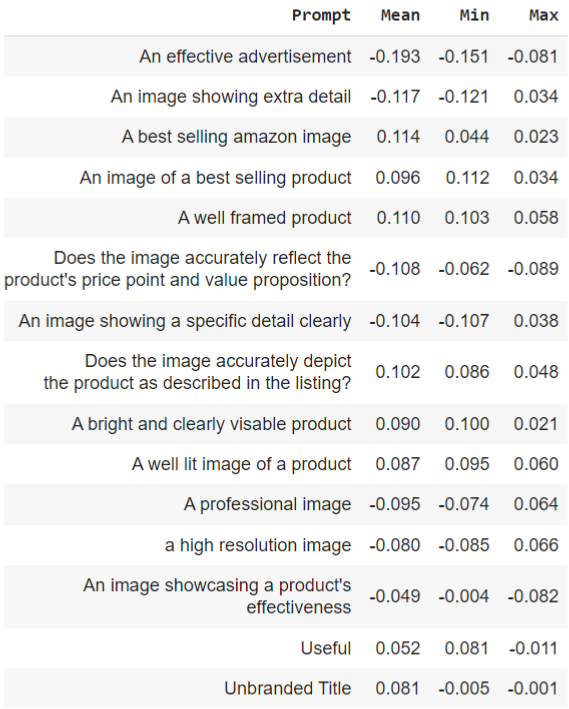


Fig. 5. Correlations between listing features, including BSR and sales metrics.



Prompt	Mean	Min	Max
An effective advertisement	-0.193	-0.151	-0.081
An image showing extra detail	-0.117	-0.121	0.034
A best selling amazon image	0.114	0.044	0.023
An image of a best selling product	0.096	0.112	0.034
A well framed product	0.110	0.103	0.058
Does the image accurately reflect the product's price point and value proposition?	-0.108	-0.062	-0.089
An image showing a specific detail clearly	-0.104	-0.107	0.038
Does the image accurately depict the product as described in the listing?	0.102	0.086	0.048
A bright and clearly visable product	0.090	0.100	0.021
A well lit image of a product	0.087	0.095	0.060
A professional image	-0.095	-0.074	0.064
a high resolution image	-0.080	-0.085	0.066
An image showcasing a product's effectiveness	-0.049	-0.004	-0.082
Useful	0.052	0.081	-0.011
Unbranded Title	0.081	-0.005	-0.001

Fig. 6. Correlations between image and prompt embedding similarities and listing BSR. Mean, Min and Max represent the mean, minimum and maximum similarity for all images in the listing, respectively.

1) Images: Research indicated that the most influential aspects of images lie in the content and semantics of the image, rather than any measurable features such as brightness or contrast. It was also found that state of the art image models are able to interpret this semantic information, therefore methods using *OpenAI*'s *CLIP* model were investigated to extract features.

The *CLIP* model has been trained on 400 million text-image pairs from the internet. It was trained to create embeddings (representations) of text and images, where the cosine similarity between a pair of embeddings is maximised, and the similarity to the rest of the data is minimised [24]. This means it is effective at interpreting what an image or piece of text is and means as an embedding; and the cosine similarity of this embedding to another will provide a score for how similar the CLIP model sees them to be.

Method's for how this could be used to quantify the effectiveness of images were then investigated. One aspect of product images known to improve their effectiveness is relevance to the product. This was investigated by calculating the cosine similarity between the product's title or name and its images. Another method involved calculating the similarity between the images and common online suggestions for how to create effective product images. Statements, questions and phrases were trialled. The correlations of these scores with the log of the products' BSRs were calculated and are shown in Figure 6.

Interestingly, both the product name and title showed positive correlations with the BSR, even after the brand name and any non-dictionary terms had been removed. This suggests that as the similarity increases, the listing's performance decreases. It is possible this is because if the images contain similar information to the product title, they do not convey useful information to the user. It could also imply that titles that are very simple or plain, so are similar to the images, aren't as eye catching to consumers. Further work could be done to investigate this effect; however the unbranded title still shows a relatively strong correlation so will be useful in the model.

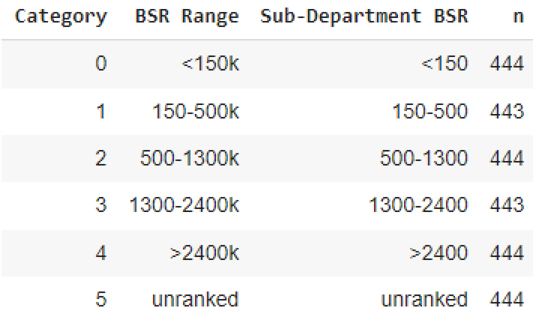
Most phrases showed a more expected negative correlation with BSR. The strongest was 'An effective advertisement', which had a relatively strong correlation of -0.19. This shows that the *CLIP* model is effective at identifying features in images which are linked to an improved BSR.

2) Product Description: Research on product descriptions has shown that, like images, the key features through which they affect a listing's performance lie in the semantics and meanings behind the description. The means that simply looking at the length, or frequency of certain words used, will not provide a truly effective way of interpreting the description's efficacy.

The product descriptions will be analysed using custom built NLP model, trained to predict the BSR of listings from their description. This model was provided with permission from [23]. The model was adapted from a larger open source pre-trained model, and then specifically trained on listing descriptions and BSRs from the same data set. It provides both a 'score' and a predicted BSR 'bin' for each description. Both metrics were trialled in the full listing model outlined in this report.

3) Best Seller Rank: The department BSR was the metric to be predicted. Different departments have different sizes, and so the same quality of product and listing will have a different BSR in another department. Most of the data provided was from the largest Amazon department - home & kitchen, and so the model was trained for these products. BSR from here on in the report will mean BSR in home & kitchen. This approach has the added advantage of greater consistency of listing features, for example the images will all be of similar objects. The importance of each feature will also be more consistent, whereas the relative importance of certain features may differ significantly between departments.

Because many of the listings had no BSR, it was treated as a categorical variable. The listings with known BSR were split into 5 even categories, and label encoded before being fed into the model. The ’unranked’ listings were then assigned another label, and a random sample equal to the size of the other categories was taken from here and added to the data. This was necessary as the number of ’unranked’ listings was far greater than the number of listings with known BSR, and this would skew the model. The category labels and their equivalent BSR ranges are shown in Figure 7.



Category	BSR Range	Sub-Department BSR	n
0	<150k	<150	444
1	150-500k	150-500	443
2	500-1300k	500-1300	444
3	1300-2400k	1300-2400	443
4	>2400k	>2400	444
5	unranked	unranked	444

Fig. 7. Assigned BSR categories, with equivalent sub-department rank and number of listings assigned.

4) Number of Images and Videos: The number of images and videos were also labelled as categorical variables, as they do not represent continuous distributions. Most listings had the maximum number of images and no videos, and there was no clear numerical relationship between these features and BSR.

5) Ratings: The average rating of a product represents how satisfied the average customer who purchased a product from the listing was. It is therefore a good metric for the actual quality of the product and will affect a listing’s BSR. While this is a measure of the products quality, rather than the listing, it would be useful to include in the model. When being used as a tool for sellers with a new product, the user could input their perceived quality of the product as their ’rating’ in order to enhance the accuracy of the prediction.

In the data, many listings have no rating. In these cases, the assigned value could either be 0 or the average rating of all products. This is related to how a customer interprets a listing that has no ratings, and so investigating which of these methods yields best results could reveal some psychology of consumers. A more detailed study could be the topic of future work.

The number of ratings is also a form of social proof which the consumer will use to determine the quality of the product. However, its relationship with the BSR is more complex. As a listing’s BSR improves it will be seen by more people, which will lead to more sales and therefore more ratings. These variables are therefore codependent. As well as this the number of ratings is not a feature that a potential user of the model would know - they do not know how well it will perform as this is what the model is trying to predict. For this reason, the number of ratings was not included as a feature of the model.

6) Other Features: The sub-department which a listing belongs to will influence its BSR in its high-level department. This is because some types of products are more popular than others. Knowing a product’s sub-department, along with knowledge of which sub-departments tend to perform better, would thus be a useful feature of the model. However, the data set at the level of sub-departments is limited. Of the 2600 listings to be used to train the model, there are 350 unique sub-departments represented, and of these 140 are seen just once. Therefore, there is not enough data per sub-department for the model to learn anything useful. A larger data set, or a data set more focused on individual sub-departments would enable this and could be the subject of further work.

The price of a product is a key feature of a listing which consumers will use to make decisions. A good price for a product is relative and closely depends on its sub-department. However, not enough data is gathered on an individual sub-department level for this type of analysis. Cheaper products do still tend to generate more sales, and therefore dominate the lower BSRs in high level departments. This explains the correlation seen in Figure 4 and indicates that price is a useful feature for the model.

The listing age may be a factor affecting the listing’s BSR. When listings are first posted, they have no BSR, this improves over time, suggesting the longer a listing is left the better it’s BSR will be. However, it can also be affected by the BSR as listings already performing well will be left up for longer by the seller. To better understand this relationship, data tracking listings’ performances over time would be needed. The model could also be improved by only including listings of a particular age range, thereby removing this variable. However, due to the complexity of this relationship and limitations on data, the listing age will not be included in the model.

7) Sales Data: Transformation methods of the numerical data in the sales model were investigated to maximise their correlations with the mean sales made. These including raising to powers, taking the logarithm and exponentiation. The transformations which lead to the highest correlations are shown in Figure 8.

For all metrics other than Rating, taking the natural logarithm led to the highest correlation. This makes sense, as the metrics span several orders of magnitude, so would be expected to follow long tailed distributions. The variations at higher magnitudes are greater in absolute terms, therefore taking the log reduces the exaggerations of these changes at greater magnitudes, creating a better fit.

Exponentiating the rating may be effective as the ratings are limited to a maximum of 5, and most listings are close to this. The exponentiated rating exaggerates the variations of ratings closer to 4, creating a better spread of data which leads to a higher correlation. These may also be useful adaptations to apply to data in the listing model.

While the decision was made to remove rating count from the listing model due to its codependence with BSR, it has been included in the sales model. This is because the rating

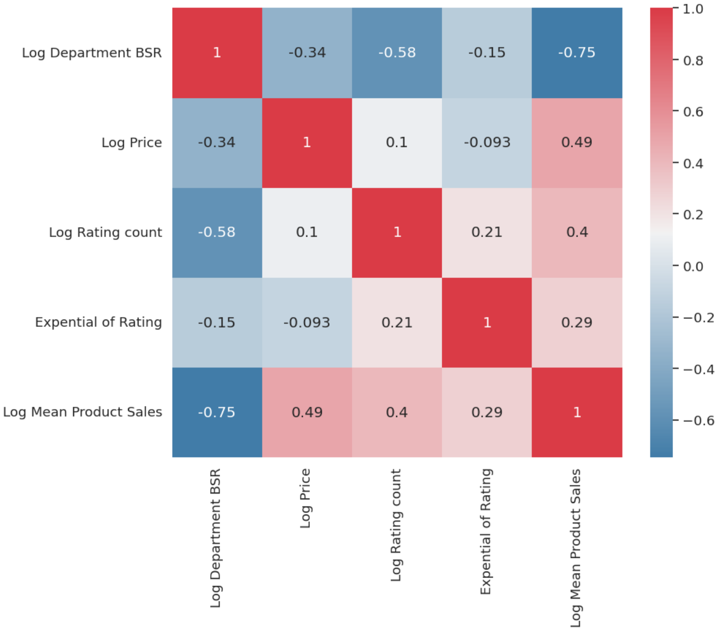


Fig. 8. Correlations between the transformed metrics for the sales model.

count is a good measure of the success of the rating, and this model is trying to measure success, rather than predict it.

B. Model Implementation

This section will detail the architectures and techniques chosen to implement the listings and sales models, as well as implementation of the user feedback tool.

1) Listings Model: The chosen architecture for the model predicting BSR from listings was a gradient booted decision tree. This was chosen as it had been proven to be effective for other listing models, was capable of predicting a categorical output and was simple to implement [8, 10]. Parameters such as the learning rate and maximum depth of the model were tuned to give optimal results. The model was trained to minimise the log loss function, which represents how inaccurate its predictions for BSR category were [25]. The output of the model is normalised using the *softmax* function by default, which converts the model's scores to probabilities representing he likelihood that the given listing belongs to each of the BSR categories [25]. The category with maximum probability represents the model's prediction.

2) Sales Model: Because all listings in the sales data had a BSR, a linear regression model could be used to predict the sales. A random forest regressor model was also trialled. The output of these models was a numerical prediction for the log of the mean sales. These models were trained to minimise the mean squared error and were compared by their R^2 values. R^2 represents the proportion of the variance in the log of sales which is explained by the model [26]. Predictions were made and then analysed for many different train/test splits, and the average taken. This removed random variations which can have significant impact on the model when using small data sets.

3) User Feedback: A basic concept was used to feedback useful information to a user of the model. They would enter their listing data into the model. The model would then output the predicted category for their listing. It would also show how metrics from their listing compare to the average listing, and the average in the best performing category. This would allow the user to understand specifically which areas of their listing could be improved to achieve to a better BSR.

V. Results and Discussion

1) BSR Prediction Model: A gradient boosted decision tree model was trained to predict the BSR category, as outlined in section IV-A. The model was trained and optimised on training and validation data sets. Its performance was then measured on hold a holdout test data set, allowing model features and hyperparameters to be optimised. These data sets were split randomly, and scores averaged over many reshuffles to remove the effect of random changes.

The best results were found when missing values of Rating were filled with the average rating value, rather than 0. This suggests that from the consumer's point of view a missing rating is seen as a lack of information about the product, rather than it being a poor product necessarily. As with the sales model, taking the logarithm of the price was found to give superior results.

The model was trained including all combinations of the top three image features' minimum, maximum, and mean values. The combination which gave the most accurate model included: the mean and minimum similarities with the prompt 'An image showing extra detail'; and the mean and maximum similarities with the prompt 'A best selling Amazon image'. This suggests that while the quality of the average image is important, the standout best and worst images also have a strong impact.

The product descriptions were analysed by the copywriting model by [23]. This gave both a 'score' and a predicted 'bin' indicating its prediction for the BSR category of the description. The predicted bin was found to be most effective in the model, likely because this represented a version of the score that had been normalised for predicting BSR.

The model's improvements in prediction accuracy, both per category and on average, as more of the data is used are shown in Figure 9. The model's accuracy is significantly higher than random guessing, implying the model is effective. The model's mean accuracy improves linearly with the amount of data, suggesting that its accuracy could continue improve significantly. The accuracy of the top and bottom BSR and unranked categories are significantly greater than the middle categories. This suggests the model is much more effective at identifying the 'best' and 'worst' listings rather than how 'good' the average ones are. This could be due to there being a large overlap in listing quality in these categories, as here outside factors such as ad spend can have a greater impact. A study investigating ad spend alongside this model may improve its accuracy. The accuracy of middle categories does increase at a faster rate than the other categories however, suggesting these categories may just require more data.

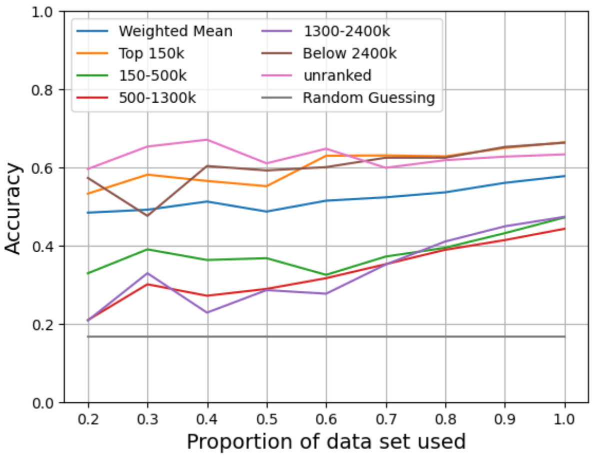


Fig. 9. Prediction accuracy for each BSR category as more of the data is used. The weighted mean accuracy using the full data set was 57%
.

The ’feature importance’ of each feature in the model is shown in Figure 10. The values represent the average gain of each feature when it was used in the decision trees, and therefore indicate the weight each feature has on predicted BSR. Rating is given the highest importance, suggesting that the better value and quality products do tend to sell more on Amazon.

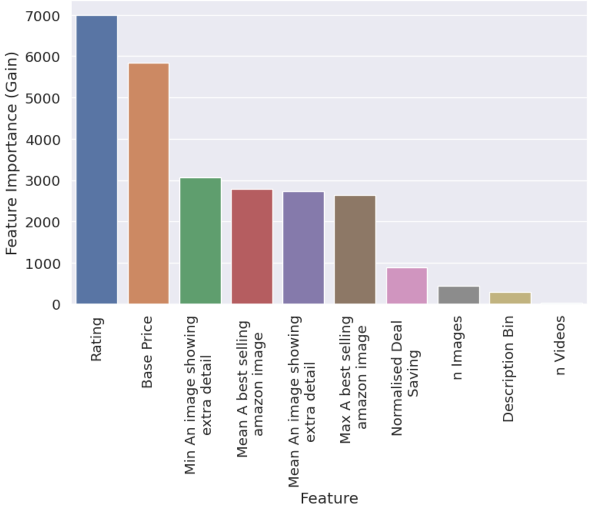


Fig. 10. Importance of each feature in the model. Importance is measured by *gain*, which represents the relative contribution of a feature over all trees in the model.

The image features were seen to be important, suggesting that the CLIP interpretations of the images are very useful. They are rated as many times more important than the number of images, which had been identified as a useful metric in previous studies. This is still a very basic implementation using this technology, and its high importance outlines the potential efficacy of a more specialised implementation. It also demonstrates the true importance of images in product listings, which hasn’t been shown in previous literature.

The description bin feature was given a relatively low importance. This is inconsistent with research on product descriptions, which suggested the description has a significant impact on performance. This may be due to limitations on the amount of data available for training the description model. Training with more data, or with a different advanced open source NLP model would likely enable more accurate ’scores’ to be assigned to the description, and would further improve the model.

2) Sales Model: Linear regression and random forest regressor models from *scikit-learn* were trained predicting the log mean sales using the transformed sales data set. R^2 scores for both models were averaged over many iterations of reshuffling and taking test and training samples, to account for random variations in the data set. The best scores for the random forest regressor and linear regression models were 0.59 and 0.61, respectively. The improvement in R^2 for the linear regression model is shown below in Figure 11. The final gradient of this graph is not zero, implying the model can improve with more data.

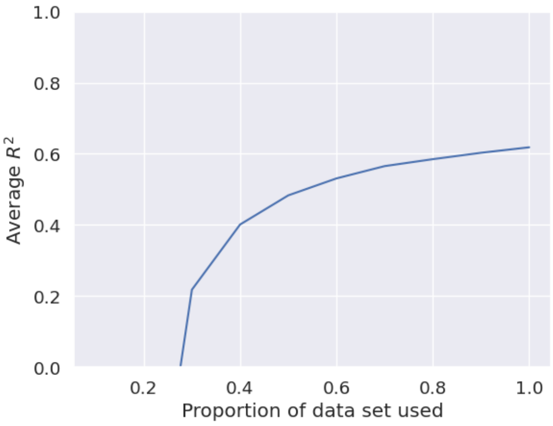


Fig. 11. Average R^2 score for linear regression model on hold out test data set. Data set was shuffled and results recalculated for 1000 iterations to account for random variation in the data. The best average R^2 score is 0.61.

The residuals for the linear regression plot are shown in Figure 12. They are seen to be randomly distributed, implying the variation is random and shows no trend, suggesting the model is a good fit to the data. The route mean squared error was 1.13. When converted from log space this equates to the average error being 3.1 times it’s predicted value. This is a good fit, given the orders of magnitude which the data spans.

Some outliers are seen in the residual plot. It is possible these are due to the model having very few data points for some departments. The model was retrained using only departments featuring more than ten times, and it’s progression curve is shown in Figure 13. The model performs much better, with an R^2 score of 0.70. 19 business reports were included in the model, and the final gradient of the curve is approximately 1.5% per business report. It was estimated that with a further 20 business reports included this would be expected to reach a maximum of 0.75. A future study could ensure data from all departments is represented evenly to build a complete model.

The best sales prediction model was predicted to be able to achieve an R^2 score of 0.75 with sufficient data. This performs moderately better than the previous study’s model, [5], which

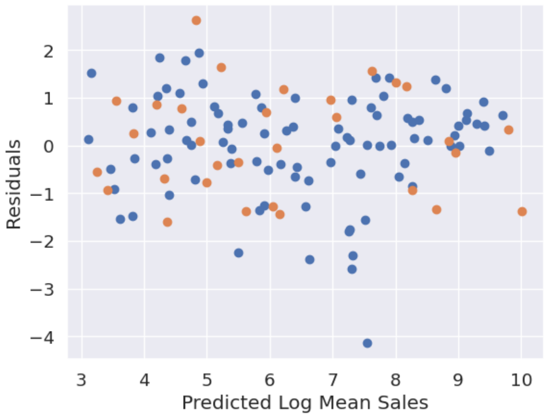


Fig. 12. Plot of the residuals for the linear regression model. They are seen to be randomly distributed, indicating the model is not biased. The route mean squared error was 1.13.

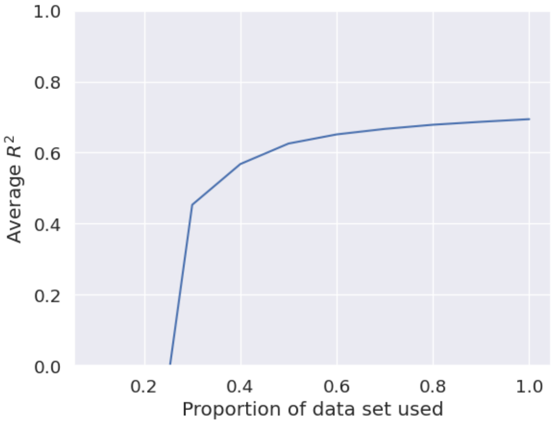


Fig. 13. Average R^2 progression when the model was trained only on data points from departments which featured more than 10 times. The best R^2 score was 0.70.

achieved a score of 0.69. This demonstrates that the inclusion of features from the listing, such as ratings and price improve the model. However, there is still 25% uncertainty remaining. This could be reduced by tracking both BSR and sales over time.

3) Deployment: A suggestion for deployment of the listing model to be used as a tool is shown in Figure 14 below. A test listing is fed into the model, which outputs a prediction for the BSR category. It also displays the graph seen in Figure 14. This shows how the test listing's features compare with the average listing, as well as one in the top category. This allows the user to determine which areas and features of their listing are strong, and which could be improved.

The graphics of the tool could be improved to better show how good the is listing for each feature. With access to other tools, it could also be improved by offering suggestions to the user for how to improve these features. For example by connecting to *OpenAI's DALL.E* the current images could be automatically edited to be more 'best selling', or a generative text model could edit and improve the description.

For this model to be deployed commercially the model's accuracy should be improved significantly. This would firstly require a larger data set, with sufficient data from each Amazon department. The advanced machine learning methods used to interpret images and text in the listings could also be further developed. The user experience with the listing tool could also be much further improved and could even use these advanced machine learning methods to generate improved listings.

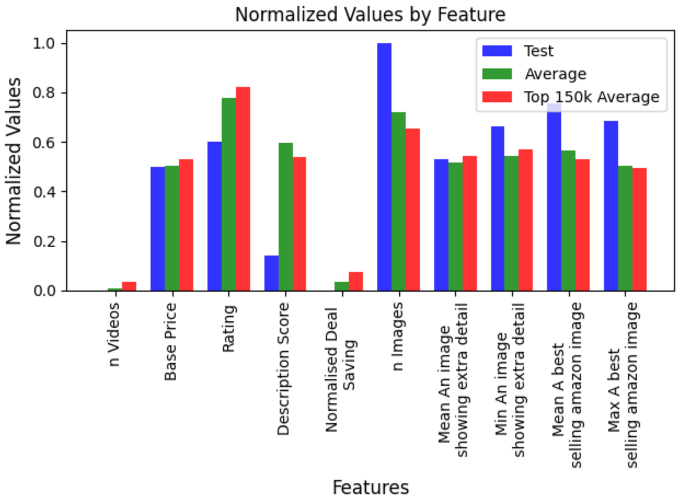


Fig. 14. User feedback for a test listing. Normalised feature values are provided alongside average and top category average to show strengths and weaknesses of the listing. The predicted BSR category was 150-500k, which was correct.

VI. Conclusions

This report has studied the relationship between features from Amazon listings and their BSR. Amazon listings had been modelled previously, however these models incorporated limited analysis of product images and description. State of the art machine learning models were used to extract and interpret semantic data from these features and used to improve a model predicting BSR. A mean BSR category prediction accuracy of 57% for the 6 categories was achieved. Image features, such as rating how similar a CLIP embedding of an image was to 'an effective advertisement', were found to be very useful features in the model. This has demonstrated the potential of the use of the latest machine learning models for semantic interpretation, and further research could refine these techniques. The copywriting model by [23] was not found to be as effective, however this was likely due to the limited amount of data available for training. Improving the model with more data or utilising another open source machine learning model could provide similar results to the images.

The relationship between BSR and product sales was also investigated. The report found that including metrics from the listing in a sales model, such as price, average rating and number of ratings, improved the accuracy of prediction - improving the previous study's R^2 score from 0.69 to 0.75. It is likely that the rest of the variation in sales is time dependent. Therefore, it is suggested that further work in this area could involve tracking the BSR and sales over time and training a model using this data.

Finally, a basic implementation using the listing model to create a 'tool' to be used by sellers was demonstrated. The tool provided the user with an estimate for a listing's BSR, as well as information on the strengths and weaknesses of their listing. Including the generated image and description features in the tool allowed it to give more detailed and specific information to the user. Requirements and suggestions for potential deployment were also detailed. This would include training the model on more data, advancing the machine learning techniques used in the model, improving graphics for the tool, and improving functionality by using generative AI to implement suggestions and improvements.

This report has shown the potential of the latest advancements in AI for predicting the performance of listings in the Amazon marketplace, using them in novel ways. It has also provided suggestions for how these methods could be further improved to create a more accurate model. Finally, it has shown how this improved model could be used to create an extremely useful tool for Amazon sellers to improve their listings.

References

[1] Lebow, S. (2022). Amazon will capture nearly 40% of the US ecommerce market. Insider Intelligence. Available at: https://www.insiderintelligence.com/content/Amazon-us-ecommerce-market

[2] Boice, M. (2022). 15 Amazon Statistics You Should Know in 2022. [online] Jungle Scout. Available at: https://www.junglescout.com/blog/Amazon-statistics [Accessed 12 Mar. 2023]

[3] ReB., & MaioN. (2020). How Amazon's E-Commerce Works?. International Journal of Technology for Business, 2(1), 8-13. Available at: https://nstudy.co/index.php/journal/article/view/136

[4] Sellerapp (2023). SellerApp: Amazon All In One Seller Tool For PPC, Keywords & Profit. [online] Available at: https://sellerapp.com [Accessed 16 Mar. 2023].

[5] A. Sharma, H. Liu and H. Liu. (2020). "Best Seller Rank (BSR) to Sales: An empirical look at Amazon.com," 2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Macau, China Available at: https://ieeexplore.ieee.org/document/9282620

[6] Li, L. (Ivy), Tadelis, S., & Zhou, X. (2020). Buying reputation as a signal of quality: Evidence from an online marketplace. The RAND Journal of Economics, 51(4). Available at: http://www.jstor.org/stable/45380744

[7] Chong, A. Y. L., Li, B., Ngai, E. W., Ch'ng, E., & Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies. International Journal of Operations and Production Management, 36(4), Available at: https://doi.org/10.1108/IJOPM-03-2015-0151

[8] Knape, D.A. (2020). Cooperation Strategies for the Amazon Marketplace. Universität St. Gallen. Thesis. Available at: https://www.alexandria.unisg.ch/publications/262425

[9] Chaudhuri, A. (2018). A Smart System for Selection of Optimal Product Images in E-Commerce. Available at: https://arxiv.org/pdf/1811.07996

[10] Kranzlein, M. (2018). A Multiple Classifier System for Predicting BestSelling Amazon Products. Thesis, Kennesaw State Universty. Available at: https://digitalcommons.kennesaw.edu/cs_etd/18?

[11] Jungle Scout (2023). Jungle Scout: Amazon Seller Software & Product Research Tools for FBA and eCommerce Businesses. [online] Jungle Scout. Available at: https://junglescout.com [Accessed 27 Mar. 2023].

[12] Pryzant, R., Chung, Y.-J. and Jurafsky, D. (2017). Predicting Sales from the Language of Product Descriptions. ACM Reference, 10. Available at: https://nlp.stanford.edu/pubs/pryzant2017sigir.pdf

[13] Openai (2023). GPT-4 Technical Report. Available at: https://cdn.openai.com/papers/gpt-4.pdf

[14] Kim, M. and Lennon, S. (2008). The effects of visual and verbal information on attitudes and purchase intentions in internet shopping. Psychology and Marketing, 25(2). Available at: https://onlinelibrary.wiley.com/doi/10.1002/mar.20204

[15] Wei Di, Neel Sundaresan, Robinson Piramuthu, and Anurag Bhardwaj. 2014. Is a picture really worth a thousand words? - on the role of images in e-commerce. In Proceedings of the 7th ACM international conference on Web search and data mining (WSDM '14). Association for Computing Machinery, New York, NY, USA, 633–642. Avaiable at: https://doi.org/10.1145/2556195.2556226

[16] Cheng, H. et al. (2012). Multimedia Features for Click Prediction of New Ads in Display Advertising. Available at: https://maths-people.anu.edu.au/~johnm/courses/mathdm/talks/dimitri-clickadvert.pdf

[17] Zakrewsky, S., Aryafar, K. and Shokoufandeh, A. (2016). Item Popularity Prediction in E-commerce Using Image Quality Feature Vectors. arXiv:1605.03663 [cs]. Available at: https://arxiv.org/abs/1605.03663

[18] Wulff, D.U., Hills, T.T. and Hertwig, R. (2014). Online Product Reviews and the Description-Experience Gap. Journal of Behavioral Decision Making, 28(3). Avaiable at: https://doi.org/10.1002/BDM.1841

[19] Chen, Pei-Yu and Wu, Shin-yi, (2007). Does Collaborative Filtering Technology Impact Sales? Empirical Evidence from Amazon.Com. Available at: http://dx.doi.org/10.2139/ssrn.1002698

[20] Bao, T. and Chang, T.-L.S. (2014). Why Amazon uses both the New York Times Best Seller List and customer reviews: An empirical study of multiplier effects on product sales from multiple earned media. Decision Support Systems. Available at: https://cyberleninka.org/article/n/1084036$

[21] Sharma, S.K., Chakraborti, S. and Jha, T. (2019). Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach. Information Systems and e-Business Management, 17(2). Available at: https://ideas.repec.org/a/spr/infsem/v17y2019i2d10.1007_s10257-019-00438-3.html

[22] Hu, N. (2021). Ratings Lead you to the Product, Reviews Help you Clinch it? The Dynamics and Impact of Online Review Sentiments on Products Dynamics and Impact of Online Review Sentiments on Products Sales Sales. Available at: https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=4508&context=lkcsb_research

[23] Towler, D. (2023). Analysing Copywriting Data from Amazon Listings. MENG Research and Developement Project, Department of Engineering, Durham University.

[24] Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. Available at: https://arxiv.org/abs/2103.00020

[25] Pedregosa F et al. (2011). Scikit-learn: Machine Learning in Python. 12(85):28252830 https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[26] Newcastle University (n.d.). Numeracy, Maths and Statistics - Academic Skills Kit. [online] www.ncl.ac.uk. Available at: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html