**MIS 637 Data Analytics and Machine Learning**

**Section E, Spring 2024**

Amazon Best Sellers Project

Report by: Aarushi Talati

Table of Contents

[Background: 1](#_Toc166102425)

[Problem Statement: 2](#_Toc166102426)

[Phase 1: Business Understanding 3](#_Toc166102427)

[Phase 2: Data Understanding 4](#_Toc166102428)

[Phase 3: Data Preparation 5](#_Toc166102429)

[Phase 4: Modeling 6](#_Toc166102430)

[Phase 5: Evaluation 7](#_Toc166102431)

[Conclusions and Recommendations 8](#_Toc166102432)

# Background:

Amazon.com is a Seattle, Washington–based e-commerce and cloud computing giant whose humble beginnings can be traced to founder Jeff Bezos’s garage, where he began selling books on the still-emerging World Wide Web. Amazon began as an online book seller, but the company went on to become a major disruptor across sectors and industries. And like many classic “disruptors” throughout history, affecting wholesale change required that Amazon invent (and sometimes acquire) new technologies and new ways of doing things.

As part of its goal to create the biggest online shopping site, in 2000, Amazon opened its store to third-party sellers, giving independent retailers a space to offer a variety of new, used, refurbished, and/or unique items. Amazon would host the platform and handle much of the logistical legwork in exchange for transaction fees. The company strengthens its partnership with sellers by offering new products and services. Some of those include incorporating warehouse fulfillment and distribution, delivery services, shipping, and inventory management, as well as customer relationship management (CRM) services into their partnerships.

Often entrepreneurs or third party sellers pick up quick selling and high demand products in wholesale and sell then individually on Amazon as a part of an easy business model. Most of the times, these products are drop shipped by Amazon. Drop shipping is a retail fulfillment method where a store doesn't keep the products it sells in stock. Instead, when a store sells a product using the drop shipping model, it purchases the item from a third party and has it shipped directly to the customer. Almost any type of product can be drop shipped, but certain categories lend themselves particularly well to this business model.

Identifying products that can become best sellers on Amazon and are likely to have a higher demand can help retailers to take advantage and increase or make inventory for such products.

# Problem Statement:

With millions of products and millions of customers, it's no wonder that retailers everywhere are trying to figure out how to tap into Amazon's market. This project will try to make a model to identify a best seller on Amazon with higher profitability margins. There are ten major factors that make a product a best seller on Amazon.

1. High Demand: Best-selling products typically fulfill a significant consumer demand within their market segment.
2. Competitive Pricing: Best sellers often have competitive prices that attract value-conscious shoppers.
3. Effective Marketing and Promotion: Successful marketing and promotion efforts play a crucial role in making a product a best seller on Amazon.
4. Positive Reviews and Ratings: Best-selling products typically have a high number of positive reviews and ratings from satisfied customers.
5. Optimized Product Listings: Best sellers have well-optimized product listings with clear, compelling, and informative content.
6. Fast and Reliable Fulfillment: Products that offer fast and reliable fulfillment options, such as Amazon Prime or Fulfillment by Amazon (FBA), tend to perform better on Amazon.
7. Seasonal Relevance: Some best-selling products capitalize on seasonal trends, holidays, or special events.
8. Product Differentiation: Best sellers often stand out from competitors by offering unique features, designs, or value propositions.
9. Continuous Optimization: Successful sellers continually optimize their product listings, pricing strategies, marketing efforts, and inventory management to adapt to changing market conditions, consumer preferences, and competitive dynamics.
10. Brand Reputation: Established brands with a strong reputation for quality, reliability, and customer satisfaction have an advantage in becoming best sellers on Amazon.

To make our model we will be focusing on the competition, pricing and reviews and ratings. Additionally, we can also classify the product listing.

# Phase 1: Business Understanding

We will be scraping and analyzing data of the products in the Computers & Accessories category from the Amazon Best Seller site. The same process can be used to scrape data from any other category on this site. Scraping data from the Amazon Best Sellers list can give you an idea of which products are currently in high demand, what features are popular among customers, what the average price range for different types of products is etc.

The attributes or features we intend to gather from the website are:

* Product Url : The link to the product.
* Ranking : The ranking of the product within the overall list of best-selling products in the category of Computers & Accessories on Amazon.
* Product Name : The name of the product.
* Brand : The brand name of the product.
* Price ( in Dollars ): Price of the product in Dollars.
* Number of Ratings : Number of ratings the product has got.
* Star Rating : The star rating the product has got.
* Size : Size of the product.
* Color : Color of the product.
* Hardware Interface : The hardware interface of the product.
* Compatible Devices : Other devices that are compatible with the product.
* Connectivity Technology : The technology using which the product can be connected.
* Connector Type : The type of the connector.
* Data Transfer Rate : The rate at which the product transfers the data.
* Mounting Type : The method to attach the product.
* Special Features : Any additional feature that the product has.
* Date First Available : The date when the product was first made available for purchase on Amazon.

The scraping will use various python libraries such as time, BeautifulSoup, random, pandas, and Selenium.

# Phase 2: Data Understanding

The ideal way to perform predictive analysis on Amazon Bestsellers would be to scrape live data from their website using one of the various methods like web scraping or using an APIs. However, to test the accuracy of our training and testing we need to use a static dataset, thus the dataset used is a sample of thousand best seller’s records. To perform basic training on our model we are using a local dataset – best sellers from amazon. The attributes are as follows:

|  |  |  |
| --- | --- | --- |
| Feature Label | Description | Data Type |
| SKU | A unique identifier code for each product. | Object (String) |
| Name | The name, brand, and model of the product. | Object (String) |
| Review\_avg | The average review rating given on a scale of 0 to 5. | Numeric |
| Review\_count | The total number of reviews on the product. | Numeric |
| Ranks | A string of all the different categories the product has a ranking in with the rank number. | Object (String) |
| Min\_rank | The highest rank for the product across all categories. | Numeric |
| Max\_rank | The lowest rank for the product across all categories. | Numeric |
| Ranks\_count | Total number of categories it has ranked in. | Numeric |
| Offer | The cost of the product offered. | Numeric |
| Offers | The number of sellers offering the product. | Numeric |
| Tag1 | Tags related to the product. | Object (String) |
| Tag2 | Tags related to the product. | Object (String) |
| Request\_date | The date the product was listed onto Amazon. | Object (Date) |

As displayed in the Business Understanding section, we are focusing on the competition for each product which is expressed as ‘offers’ feature in our dataset. The pricing of the product is valued by the ‘offer’ feature whereas ‘review\_avg’ and ‘review\_count’ describe the consumer feedback for each product. Lastly, our target variable is the ‘min\_rank’ since this prediction will help sellers gauge the consumer response to the products. It would also be interesting to evaluate the predicted ‘min\_rank’ versus ‘max\_rank’ and the distance between the two to establish which products are better and safer investments.

# Phase 3: Data Preparation

The initial dataset had 13 columns with multiple missing values, non-numeric data types, varied scales and required layers of cleaning followed by normalization. Thus, this phase is divided into five sub-phases as follows:

For phase one, cleaning the data, we dropped all the string columns including ‘SKU’, ‘Name’ and ‘Ranks’ along with columns that had more than seventy percent missing values like ‘tag1’ and ‘tag2’. For the other columns with missing values, we perform linear regression to estimate the missing value and fill it in. After performing this step, all features were manipulated to be numeric and complete with no null values. Before normalizing the data, we separate the target variable from the features and store it as y and X respectively.

Normalization rescales the features of a dataset to a standard range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. This process ensures that all features contribute equally to model training, preventing larger-scale features from dominating the learning process. By normalizing data, we facilitate faster convergence in optimization algorithms, prevent regularization biases, and ensure accurate distance calculations in distance-based algorithms. Normalization enhances model performance, stability, and interpretability, making it a fundamental preprocessing step in machine learning workflows.

Now, we can go ahead and scale all features in X using z-score normalization, so our dataset is ready to be fit to our model. We are using z-score normalization here since it is more robust to outliers, easy to interpret when the mean is zero and the distribution of the shape of the feature is maintained. However, when we evaluate our models and try the neural network, we will reprocess the data to range from 0 to 1 instead by using min max normalization.

The last step before we start to train our model is splitting our data into appropriate ratios of training and testing datasets. The standard ratio is eighty to twenty percent, which is also what we are going to use.

# Phase 4: Modeling

**A screenshot of a graph

Description automatically generated**As preliminary analysis, we plotted a correlation heatmap to find out which features are significantly related to the target variable and concluded that the maximum rank and number if ranks is highly correlated to the top rank whereas other features play an important role to define these three attributes of the model.

Our goal is to effectively predict the ranking of a product to allow users of our model to forecast which products are investible and have fewer sellers, which means less competition.

Since this is supervised predictive modeling there were a lot of model choices to weigh and choose from. The most basic modelling technique would be Linear Regression which is easy to interpret and efficient for processing large sets of data. However, feature relationships tend to be more complex and non-linear in nature which this modelling technique cannot capture. On the other hand, we have the K-Nearest Neighbor technique which is effective for prototyping and baseline modelling. Since it is non-parametric, it does not make strong assumptions making it more versatile. The drawback is that it tends to be more sensitive to outliers and is expensive to process larger datasets. Decision trees are comprehensive and easy to visualize while capturing complex and non-linear relationships. Although for our dataset we require prediction of a continuous numeric data type which is why this might not be the best choice. Also, decision trees are prone to overfitting and are not effective for biased datasets like this use case.

Another powerful method is neural networking which is flexible and scalable. It can learn complex nonlinear relationships but is also complex to understand because of its black box nature. Lastly, K-Means clustering is a unique method that can identify underlying patterns but no target prediction. This makes it an appropriate choice for a preliminary model to understand the trends in the dataset and then make a supportive model choice. However, the initialization will strongly affect this model and its results, which need to be taken into consideration.

The table below weighs the merits and drawbacks of the multiple model options.

|  |  |  |
| --- | --- | --- |
| **Model Type** | **Merits** | **Drawbacks** |
| Linear Regression | Easy to interpret.  Efficient for large sets of data. | Feature relationships are more complex and nonlinear. |
| K-Nearest Neighbors | Easy for prototyping and baseline modeling.  Non-parametric thus versatile. | Expensive for large datasets.  High sensitivity to noise and outliers. |
| Decision Tree | Captures complex non-linear relationships.  Easy to visualize. | Apt for binary predictions or classification.  Prone to overfitting. |
| Neural Networks | Flexible and Scalable.  Can learn complex non-linear relationships. | Extremely complex.  Black box nature. |
| K-Means Clustering | Apt for identifying underlying patterns. | No target prediction.  Sensitive to initialization. |

# Phase 5: Evaluation

Even though practically Modelling and Evaluation go iteratively, we will discuss then chronologically here. All the above models were evaluated using a series of different metrics like accuracy, precision, mean squared error (MSE) and root mean squared error (RMSE).

Taking the above into account, we proceed to build a K Means clustering first along with a scatter chart to visualize the same.

A graph of colored dots

Description automatically generated with medium confidencePCA component 1 refers to the linear combination of all original features and explains the largest possible variance against the target variable. This proves correlativity of all the features to ‘min\_rank’.

We can see the underlying patterns and clusters from this graph.

A graph with blue dots

Description automatically generatedNext, we try a linear regression and K-Nearest Neighbor model as a baseline model for the dataset.

The linear regression shows a lot of variances between the predicted and actual proving that the target variable and the features don’t share linear relationships.

The simple KNN model with five neighbors gave satisfactory results with a Mean Square Error (MSE) of 141 and testing accuracy of eighty three percent with was not too variant from the training accuracy of eighty nine percent. Upon performing cross validation to find the best parameters, eight is the number of nearest neighbors that gives a mean square error of 394.

Lastly, we tried a basic neural network with an input, three hidden layers and the output. The hidden layers had 128, 64 and 32 nodes respectively. The model produced a training mean-squared error of 157 and testing mean-squared error of 205 which are relatively high. Thus, the neural network requires a lot of fine tuning to be deployed.

# Conclusions and Recommendations

Since the dataset was a sample dataset, we would require a larger and more diverse data set to train our model more effectively. Real-time data would be highly recommendable for this project, using techniques like web scraping. We can also explore category and region-oriented models as those features strongly define the bestsellers. More advanced techniques of sentence embeddings for sentiment analysis and image embeddings for product reviews can be used to refine the model.

**References:**

<https://www.britannica.com/topic/Amazoncom#ref371216>

<https://www.blog.datahut.co/post/scraping-amazon-best-seller-data-using-python>

<https://medium.com/@MinatoNamikaze02/running-selenium-on-google-colab-a118d10ca5f8>