Final Report

**Abstract**

This study presents an exploration of Amazon co-purchasing data with the primary aim of uncovering hidden patterns and insights through the application of machine learning algorithms. The direct problem at hand involves the analysis of Amazon co-purchasing data, with a particular emphasis on leveraging clustering algorithms to identify meaningful patterns within the dataset. My personal objective of this research is to develop a decisive algorithm and an accompanying application that not only facilitates a comprehensive understanding of Amazon's co-purchasing data but also enables the discovery of latent connections and associations among the data points. By utilizing clustering algorithms, this application seeks to categorize products and user behaviors into distinct groups, thus shedding light on purchasing trends, preferences, and potentially revealing novel insights for both customers and sellers on the Amazon platform. This endeavor aims to contribute to the field of e-commerce data analysis by providing a practical tool for deciphering co-purchasing data, ultimately assisting in the optimization of product recommendations, marketing strategies, and user experiences on the Amazon platform.

Introduction:

This project provided me a strong understanding of building multiple programs to further understand the data that was provided by the SNAP Stanford data surrounding Amazon Meta Data from 2006. My plan was to first decipher the data by organizing it into an easy, readable format using a .json build. After the data was easily readable, I started to build multiple programs to gather information that would further help me understand what the data was able to tell me and how the co-purchasing data could provide further insight on how Amazon suggests extra purchasing. Using machine learning algorithms, I was able to build applications that would tell me the Top 50 of a specified group, applications that could predict what I would want to buy based on reviews or average Amazon rank, as well as build a predictive User Interface that would suggest to me the Top 5 of a selected group or predict a suggested item using mean square calculations.

Background and Motivation:

This project represented a significant step in my journey to comprehensively analyze and extract meaningful insights from complex datasets. Focusing on the SNAP Stanford dataset containing Amazon Meta Data from 2006, my primary objective was to unravel the underlying trends and patterns within this rich source of information. To achieve this, I devised a systematic approach, beginning with the transformation of the raw data into a structured and easily readable .json format.

Once the data was organized, I embarked on a multifaceted journey of program development. My aim was to harness the data's potential fully and gain a profound understanding of Amazon's co-purchasing dynamics. Utilizing machine learning algorithms, I created applications that could reveal valuable information. These applications ranged from identifying the Top 50 products within specific categories to predicting personalized purchase recommendations based on factors such as product reviews and Amazon rank averages.

Moreover, I designed a predictive user interface that could offer me the top 5 product suggestions from a selected group, enhancing the shopping experience and streamlining decision-making on the Amazon platform. This project not only empowered me with a deeper understanding of data analysis and machine learning but also equipped me with practical tools to navigate the complexities of e-commerce data, ultimately facilitating more informed purchasing decisions and user experiences.

Approach:

1. My approach to tackling the complexities of the SNAP Stanford Amazon Meta Data from 2006 involved a systematic and structured workflow. The project unfolded in several key phases, each contributing to a comprehensive understanding of the data and the development of valuable tools for e-commerce analysis.
2. First and foremost, I began by addressing data organization. The initial step involved transforming the unstructured .txt data into a more manageable and structured .json format. This critical data preprocessing stage laid the foundation for subsequent analysis.
3. With the data in a more accessible format, I proceeded to implement machine learning programs to extract insights. I categorized the data into four distinct groups, facilitating a more granular analysis of Amazon's co-purchasing patterns. This segmentation enabled me to discern trends and behaviors within each group.
4. Next, I developed tools to provide actionable recommendations. I created algorithms capable of identifying the Top 50 product suggestions within each group, along with related items, offering users a comprehensive view of their options. Additionally, I implemented a basic search function that allowed users to receive product suggestions based on specific categories, further enhancing the customization of recommendations.
5. To take the analysis a step further, I designed a predictive learning model. This model utilized machine learning techniques to predict user preferences within each group, aiding in the generation of personalized product recommendations based on reviews and mean square calculations of Amazon rank. This predictive capability added a dynamic dimension to the recommendation system.
6. Lastly, I developed a user-friendly interface that not only presented the top 5 product suggestions from a selected group but also harnessed the predictive power of the model to provide users with additional recommendations based on user preferences. This interface streamlined the decision-making process, offering a holistic shopping experience on the Amazon platform.
7. In summary, my approach encompassed data preprocessing, segmentation, recommendation generation, predictive modeling, and user interface development. This systematic workflow enabled me to harness the power of machine learning and data analysis to decode Amazon's co-purchasing data, offering valuable insights and enhancing the user's shopping journey.
8. ***Evaluation***

Once I have completed the work for the file organization, I was able to separate the data into readable, functional data that could provide insight into Amazon’s Co-Purchasing data.

The first set of applications I made gave me insight into what books, dvds, music, and videos were considered popular as well as highly reviewed. In an attached excel files (top\_books.xlxs, top\_dvds.xlxs, top\_music.xlxs, and top\_videos.xlxs), you will see that the data will show favored items based on highly rated reviews, as well as Amazon Rank. From there, I noticed that the data provided related suggestions that were either considered to be similar in reviews, similar in Amazon Rank, or considered to be similar via the category that item was in. Some of the data will show basic ASIN (Amazon Serial Identification Number) due to the data being “cherry picked” for the basic use in a school function. This data did provide insight into how Amazon provides a suggested Co-Purchase idea by simply adding something identical to what you were buying. This is a basic process, so I had to create other applications to dig a little deeper.

The next application I built was centered around providing me an item to buy based on category. The idea behind this build was that I would ask for the application to run through the data and give me the best possible option an item with the category listed vs Amazon rank. When diving into the business analytics of Amazon, I tried to design something that would simulate a “make a decision for me,” option to see how the data would react to what I should buy. This application did give me the result of suggesting a co-purchase item that would corollate to what I am buying at that time. The data proved that category was in fact important, but not as important as reviews and Amazon rank.

After building the previous applications, I needed to build something that would predict what anyone would buy based on a predicted sales rank. The question I asked was, “if someone were to buy something, what are the odds that they would buy a specific title, and would they want to buy something with that?” In the attached excel files (predicted\_books.xlxs, predicted\_dvds.xlxs, predicted\_music.xlxs, and predicted\_videos.xlxs), you will notice that based on the hyperparameter of predicted sales rank, people would most likely buy such titles as you would find in an example shown in Figure 1. What this state is that “Robin Hood – Men in Tights” would be commonly bought because so many other people have bought it before, based on an average I build in the applications.

\*Note: This is just one of the examples I am showing in Figure 1, and other attached excel files will show Books, DVDs, and Music.

Figure 1:

|  |  |
| --- | --- |
| **title** | **predicted\_salesrank** |
| Robin Hood - Men in Tights | 1698.907536 |
| Hamlet | 1698.907536 |
| The Muppet Christmas Carol | 1698.907536 |
| For Your Eyes Only | 1698.907536 |
| The Wiz | 1698.907536 |
| Live and Let Die | 1698.907536 |
| Pride and Prejudice | 1781.885 |
| Patch Adams | 2102.450464 |
| Sliding Doors | 2102.450464 |
| Patch Adams | 2102.450464 |

Lastly, I wanted to create a User Interface that would provide other examples of suggested Top 5 of a specified group, or a predicted Top 5 of a specified group. I used mean square of the Amazon Rank, and it would learn based off of the data and give an output that would give the user a suggested item to buy. I included a “Top 5,” to simulate the idea of Amazon’s Co-Purchasing implementation where the user would be happy to either buy 1 item, or multiple items based off of the results given.

After building these multiple applications to decipher the Amazon Co-Purchasing Data, I found that most co-purchasing is based on popularity of what was bought before. If you are buying a common book, then the suggested item to buy with it is something many others bundled together in previous transactions. The data also shows that if you were to buy something that is not that popular, then (in 2006) there would not be as many suggestions to bundle something together with what you intend to buy.

1. ***Conclusion and future work***

This project provided me with many potential tools to help with my future in analytics as well as Full Stack Web Development. I would assume that the work that I have done here with data based in 2006 would be considered obsolete to today’s standards but found that the information I had to go through was exciting to go through. From my personal experience in 2006, I remember when everyone in the world stated that, “Amazon will not last long because people want to hold onto what they are buying before buying it.” I was at the age of 16 when this data was gathered, and all I heard from Amazon was that they were the company that only sold books. Advertising was not what it is today with Amazon, but it has upgraded since then.

Amazon is a leading organization when it comes to suggesting something to be bundled together with what you buy today. The meta data that was provided from 2006 shows that most of what they offered was popular and generalized suggestions based on what people really liked at a given time. The future work that has been done to Amazon Co-Purchasing Data is live in 2023. Amazon had its foundations of suggesting a bundled item that was popular and morphed it into making sure that the best item to buy with what you are looking at now. They have included (what I presume) many other hyperparameters into their strategy of Co-Purchasing today that make the suggested item more logical and secure with your purchase. Amazon has plenty of Data to learn, but I also presume that we will not have the luxury of looking at the latest data only for Amazon to protect there assets.