Pham Duy Anh Vu

FINAL PROJECT REPORT

1. Introducing the Dataset:

The link: [https://www.kaggle.com/datasets/swathiunnikrishnan/amazon-consumer-](https://www.kaggle.com/datasets/swathiunnikrishnan/amazon-consumer-behaviour-dataset) behaviour-dataset

This dataset analyzes the behavioral analysis of Amazon's consumers. It consists of a comprehensive collection of customer interactions and browsing patterns within the Amazon ecosystem. The dataset captures a wide range of Gender, Customer demographics, Purchase Categories, Frequency, Shopping Satisfaction, Reviews, all providing insights into customer preferences, shopping habits, and decision-making processes on the Amazon platform.

The purpose of this project is to analyze this Amazon consumer behavior dataset to understand consumer behavior, identify trends, and optimize marketing strategies.

1. Cleaning / Processing the Data:
2. Data Cleaning:

Before the analysis process, we need to ensure that there are no missing, incorrect values, or outliers in the Amazon Customer Behavior dataset. The code “amazon.isnull().sum()” calculates how many missing values are in each variable, and we see that only the variable “Product\_Search\_Method” has missing values (2 values). However, because our analysis does not involve “Product\_Search\_Method”, it is not necessary for us to alter the dataset. As for outlier values within the dataset, because the numerical values within the Amazon Customer Behavior data only consists of “Age” (which is not necessary to be assessed for outlier values) and values with specific grading systems that already has built in max-min values

1. Data Processing:
   1. Association Rule Mining: Apriori

* Purpose:

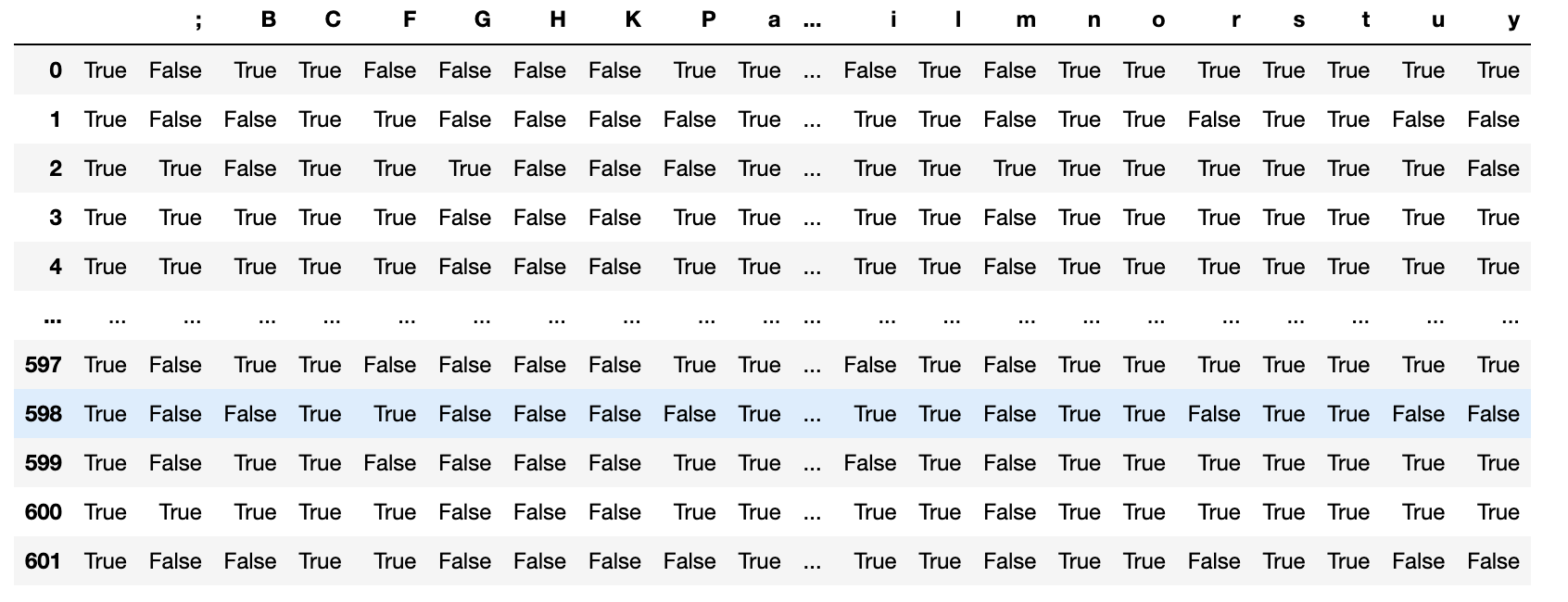
Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. Based on this, we decided to utilize the Apriori algorithm to identify frequently purchased Amazon products to understand customer buying patterns.

* Initial Data Processing Methodology:

We started of by isolating the “Purchase\_Categories” variable into a new dataset called “apriori” to process the values into values suitable for computer generation. From there, we encoded the data using the “TransactionEncoder” method from “mlxtend” package. This, however, resulted in the Apriori algorithm returning the frequency of each letter of the values within the “Purchase\_Categories” variable instead of calculating the frequency of each unique value.

* Result:

To resolve this problem, we consulted with all our available resources (from ChatGPT, DataCamp, to online resources on how to use the “mlxtend” package) and tried using different approaches including creating a dummy variable to use One-hot Encoding and even creating a mapping system to assign a numerical variable to the values. However, none of these methods worked and we ran into the same problem.



* 1. Classification: Decision Tree
* Purpose:

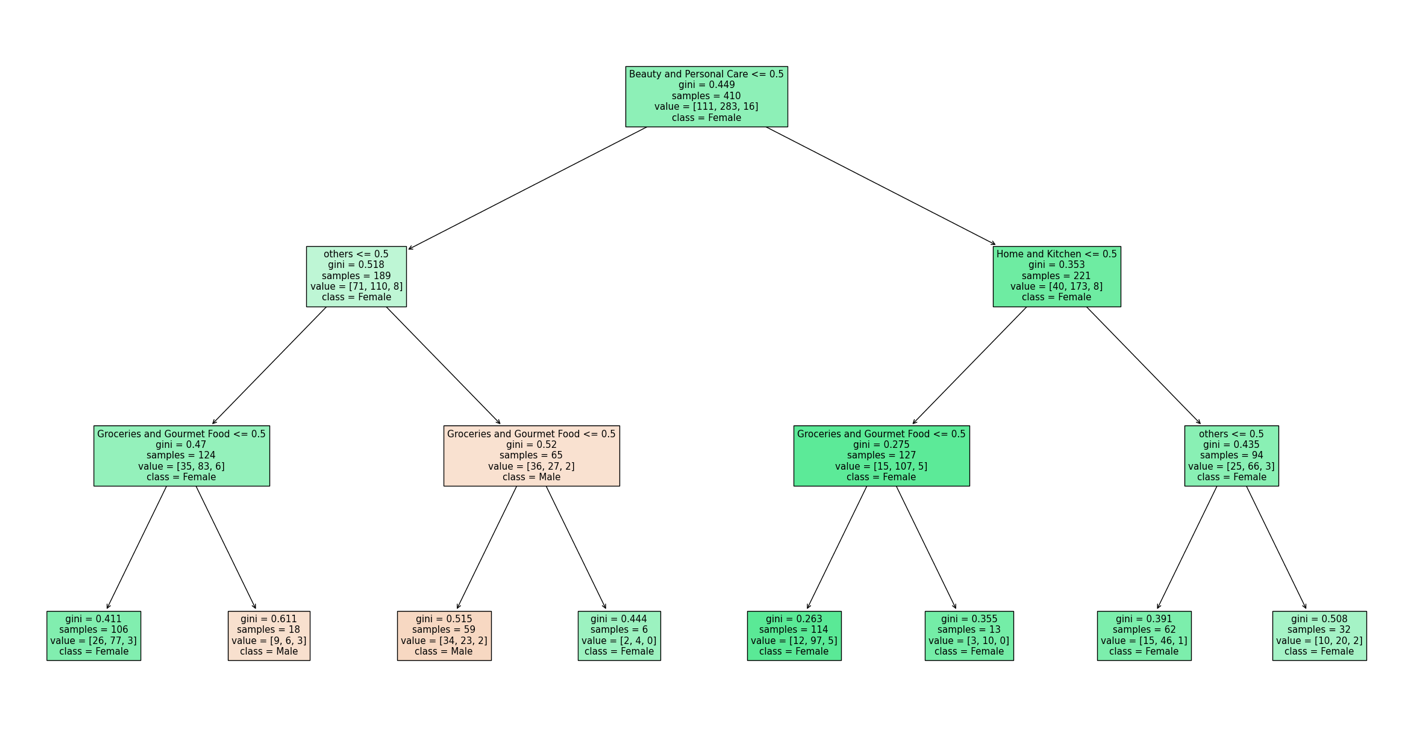
Decision tree classification is a machine learning algorithm used for predictive modeling and decision-making tasks. Its primary purpose is to create a mapping from input features to target labels by recursively partitioning the feature space into subsets, based on the values of the input features, and assigning a class label to each partition. In the Amazon Customer Behavior Analysis, we wanted to figure out the relationship between a customer’s gender and their purchase preferences and which purchasing categories each gender gravitate towards.

* Initial Data Processing Methodology:

To start with, we created a new data frame called “classification” which included the two variables “Gender” and “Purchase\_Categories”. For the “Gender” variable, we decided to remove all values of “Prefer not to say” as those values weren’t necessary for our analysis. We then created a mapping system to assign the 3 unique categorical values (Male, Female, and Other) to a numerical form suitable for processing. As for the “Purchase\_Categories” variable, we tried the method of creating a dummy variable to convert it into binary variables corresponding to each unique values in the original “Purchase\_Categories” variable.

* Result:

For the Decision Tree Classification, this method worked fine, and we achieved a result as returned by the graphed decision tree below.



The decision tree provides a model of Amazon customer behavior, with specific reference to product categories that might predict a customer's gender. Initial splits indicate the influence of both mainstream and niche categories in purchasing decisions. Subsequent branching shows a clear gender-based pattern, particularly highlighting the significance of 'Groceries and Gourmet Food', 'Beauty and Personal Care', and 'Home and Kitchen' in predicting female customer behavior, as denoted by the lower Gini scores. Interestingly, the classification of 'Male' customers is less represented in the tree; however, nodes associated with higher Gini scores and classifications of 'Male' suggest more mixed purchasing patterns. This indicates that male customers may have a broader range of interests or less distinct shopping patterns in the categories considered, which could be an area for further exploration to refine the model. Such nuanced insights are vital for Amazon to create balanced marketing strategies and to enhance recommendation systems, ensuring they cater effectively to the diverse preferences of all customer segments. Maintaining accuracy, mitigating bias, and embracing the evolving landscape of consumer behavior remain essential to the model's applicability and integrity.

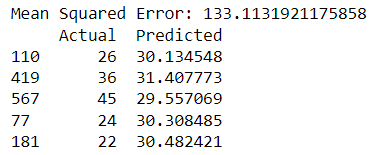
* 1. Regression: Linear Regression
* Purpose:

Linear regression is a statistical approach that models the connection between a dependent variable and one or more independent variables. Useful for forecasting and decision-making, it delineates the relationship's strength and pattern. In this instance, we are using the variables “Customer\_Reviews\_Importance” and “Shopping\_Satisfaction” to predict the variable “age”. In essence, the goal is to figure out which age group is most reliant on customer reviews to products and how helpful these reviews are, helping us figure out which age group to target for advertisement.

* Initial Data Processing Methodology:

Similar to the classification process, we create a new dataset called “regress” containing the target variable “age” and two predictors “Customer\_Reviews\_Importance” and “Shopping\_Satisfaction”. Since there is no need for further data processing, we implement the code in lab03 for linear regression directly onto the dataset to achieve the results below.

* Result:



The linear regression analysis yields a Mean Squared Error (MSE) of 133.1132, which signals a relatively high level of prediction error in estimating customer ages. This value of MSE implies that on average, the model's predictions deviate from the actual age data by a square root of roughly 133.1132, suggesting the model's predictions are not closely aligned with the actual values. It is notable, however, that the model's predicted ages cluster around the 30-year mark. This clustering effect may indicate a demographic trend within the dataset, where customers in their thirties are more consistently influenced by other customers' reviews, possibly because this age group engages more actively with reviews or because their shopping experience aligns more closely with the opinions expressed in such reviews.

In an effort to refine the precision of our linear regression model and reduce the Mean Squared Error (MSE), we tried to identify additional quantitative variables that could be integrated into the analysis. However, the current dataset did not offer further quantitative attributes that held relevance for our targeted prediction goals. We also tried to integrate a different regression method, the k-Nearest Neighbor regression method, but the results remained the same as it generated a MSE of 155.3762.

A potential solution to this problem is to merge our current dataset with another Amazon customer analysis dataset with more relevant numerical values to allow for predictors within the regression analysis.

* 1. Clustering: K-means clustering
* Purpose:

K-means clustering is a method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. In the context of the Amazon Consumer Behavior dataset, we are utilizing k-means clustering to group into three types of customers: customers with high purchasing and browsing frequency, customers with low purchasing and browsing frequency, and customers with high browsing frequency but low purchasing frequency; all of which could be used for further customer behavior analysis.

* Initial Data Processing Methodology:

We started out by encoding the categorical variables “Purchase\_Frequency” and “Browsing\_Frequency” to make it suitable for processing. We also had to scale the three variables we used for analysis as their encoding values differed in magnitude. Applying the k-means code and knowing that we wanted to divide the dataset into 3 clusters, we set k = 3 as the number of optimal clusters.

* Result:

A screenshot of a computer screen

Description automatically generatedA screenshot of a shopping survey

Description automatically generated

+ Targeted Marketing: High Frequency Shoppers might respond well to loyalty programs and new product alerts. Occasional Shoppers could be targeted with promotions and sales to increase their purchase frequency.

+ Improvement Initiatives: For Low Satisfaction Shoppers, further analysis could be conducted to understand the low scores and improve their shopping experience.

A diagram of a scatter plot of customer clusters

Description automatically generated

The 3D scatter plot visualizes the clustering of Amazon customers based on scaled metrics: purchase frequency, browsing frequency, and shopping satisfaction. Cluster 1, represented in blue, appears to be characterized by moderate to high purchase frequency and browsing frequency but spans a wider range of satisfaction levels. This indicates a segment that is actively engaging with the platform, yet their satisfaction is not consistently high, suggesting opportunities to enhance their experience. Cluster 2, in green, shows a trend of lower activity in both purchasing and browsing, and their satisfaction levels are generally around the median, which aligns with the characteristics of Occasional Shoppers who might be induced to shop more frequently with the right incentives. Lastly, Cluster 3, in red, seems to encapsulate customers with lower purchase frequencies, variable browsing frequencies, and a broader spectrum of satisfaction, possibly indicative of the Low Satisfaction Shoppers. They're engaged enough to browse but might need improvements in their shopping experience to convert browsing into purchasing. This visualization aids in understanding the relationship between customer behaviors and satisfaction, and it can serve as a foundation for tailored customer engagement strategies.

III. Conclusion

Our investigation of the Amazon Consumer Behavior dataset has provided significant new understandings into the complex web of preferences and behaviors related to online shopping. Through the application of diverse data analysis methodologies, including decision trees for classification, linear regression, k-means clustering, and Apriori for association rule mining, we have effectively discerned significant patterns and trends that can be utilized to optimize their marketing tactics and elevate customer contentment within Amazon shopping. Overall, the utilization of these analytical approaches has not only improved our comprehension of client dynamics, but it has also highlighted the significance of ongoing data monitoring and model updates. This analysis can be used to help Amazon develop proactive and reactive strategies that adjust to the changing needs of its customers. For future additions for this research project, the best method would be to incorporate additional datasets into this study to enhance the variables and reduce prediction errors, particularly in association rule mining and regression analysis. Experimenting with more complex machine learning models may also yield more accurate forecasts and deeper insights.