**Targeted Marketing Through Classification and Clustering**

**Section 1: Introduction**

Targeting customers who are likely to make a purchase is an essential goal in marketing. Making sure individuals are likely to react to an offer or advertisement is crucial, as according to recent research targeted ads are found to be 46% more effective than untargeted ads (Eser, 2024). It is also observed that the return on investment for personalized marketing efforts is five to eight times higher than non-personalized advertising (Lindner, 2024).

To attain this goal, it is necessary to make predictive models that are both comprehensive in providing a list of individuals that could be susceptible to targeted marketing, but also concise and do not cause a waste of marketing resources. Using multiple models to narrow in on individuals who are highly likely to make purchases could be a more robust method of properly allocating resources in marketing efforts. Each model used in this analysis has a condition for being considered for being targeted, and if the conditions for both models are filled, individuals would be put on the target list. These individuals on the target list would then be targeted with additional offers or advertisement, as they are more likely to be susceptible to targeted marketing due to their similar characteristics to those who accepted an offer from previous marketing efforts. To complete this goal, this analysis employs the use of principal component analysis, K-nearest neighbor classification, and hierarchical agglomerative clustering using retail marketing data that displays if individuals accepted a marketing effort directed toward them.

**Section 2: Methodologies Used**

**Data Set Description:** The data set used in this analysis is a marketing data set that contains various purchasing and demographic characteristics on customers being targeted through a marketing effort. The acceptance of a marketing offer was the target variable in this analysis, as if the individual accepted the marketing offer to purchase a product from the company they are represented by a 1, and if not, they are represented by a 0. The goal is to identify individuals that have not made a purchase from the marketing offer that are likely to with additional targeted marketing. There is also information on many different purchasing habits of the individuals, such as the different kinds of foods they have bought in the last two years as well as the number of times they visited the company’s website in the last two years. Additionally, it contains information on the individual such as level of education, number of small children and teenagers, marital status, date of enrollment with the company, if the customer complained at all during the two-year duration, and other demographic characteristics. There are 2,205 observations in total from the data set.

**Principal Component Analysis:** In this analysis, principal component analysis (PCA) is used to deal with dimension reduction and possible redundancy issues in the data. PCA is a method of changing the data where a large set of variables is reduced to a smaller number of variables to avoid possible issues in the data such as correlation and redundancy. It can capture most of the variance in the data while remedying these issues, as the PCA used in the models still accounts for 95% of cumulative variance while still reducing the number of variables used. The data is standardized before implementing PCA to make sure that individual variables did not have a disproportionate effect in the recalculation of the variables.

**K-Nearest Neighbor Classification:** The main model used for targeting possible customers in this analysis is the K-Nearest Neighbor’s model (KNN), which classifies customers based off their characteristics as accepting of the marketing offer or not. The data was split into a training and validation set for this methodology, where 60% of the observations, or 1323 observations were used for training, and 40%, or 882 observations were used for validation. KNN calculates the Euclidian distance between different observations and then compares it to other similar observations within the training set. After it is compared to the “K” number of nearest neighbors, it is then classified as an individual who is expected to accept the offer or to reject the offer. Those expected to accept the offer but did not actually accept the offer fulfill one of the two conditions necessary to be added to the target list.

**Agglomerative Clustering Support Model:** The support model in this analysis that acts as an additional filter to narrow down the list of targeted customers is a form of hierarchical clustering known as agglomerative clustering. In this clustering method, each observation starts as its own cluster, and over time begins to join with similar observations to create fewer overall clusters that contain more values. In hierarchical clustering, the process continues until all observations merge into a single cluster, known as the root. However, the merging process is typically stopped before reaching this final cluster, once the desired number of clusters is achieved. Ten clusters are chosen in this study, and the top two clusters in terms of acceptance rates are then used as a condition for making into the final target list. The final target list is a list of individuals who fulfill both the classification and false positive conditions, and these are individuals who are worth going back and marketing towards.

**Section 3: Results**

The correlogram indicated that there was a high level of correlation between many variables, justifying the use of PCA as the data set contained 33 variables. PCA indicated that the amount of PCA variables required to account for 95% or more of cumulative variance was 22 PCA terms or variables as seen in Figure 1. The PCA variables were then used in the KNN analysis to predict whether the customers are likely to accept the offer. The KNN results indicated that a k value of 9 was optimal (Figure 2) meaning that each observation is compared to its nine nearest neighbors, and the model had an accuracy of 0.8237843. The model was fairly accurate in its predictions; however, the goal of the model is not necessarily to be accurate, it is to identify false positives. The criteria for individuals to be added to the list was if they are false positive, as the KNN model believes that they have characteristics that are similar to those who did accept the offer. There were 34 false

positives, or individuals that passed the first condition for the target list as seen below:

|  |  |  |  |
| --- | --- | --- | --- |
| True Positives | False Positives | True Negatives | False Negatives |
| 60 | 34 | 654 | 134 |

The model did have some issues identifying which values are positive, as the precision, or number of observations that were identified as positive actually being positive, was relatively low at 63.8%. This inaccuracy indicates that a second model could be helpful to narrow down the target list to ensure similarity to other positives.

The agglomerative clustering model was then used to further filter down the number of individuals on the target list. The agglomerative clustering model had a very high clustering coefficient of 0.9857511, indicating that the model clustered the data points effectively (Figure 4). The data points were then broken up into ten clusters, where the two clusters with the highest rate of positive responses to the offer were chosen as high frequency clusters. The breakdown of the clusters containing **the percentage of individuals who accepted an offer** and the **total number of individuals in each cluster** is seen below:

**Cluster Statistics**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| 39.6% | 9.6% | 9.1% | 21.1% | 17.9% | 12.5% | 40.5% | 13.4% | 29.2% | 0% |
| 192 | 156 | 132 | 152 | 56 | 24 | 74 | 67 | 24 | 5 |

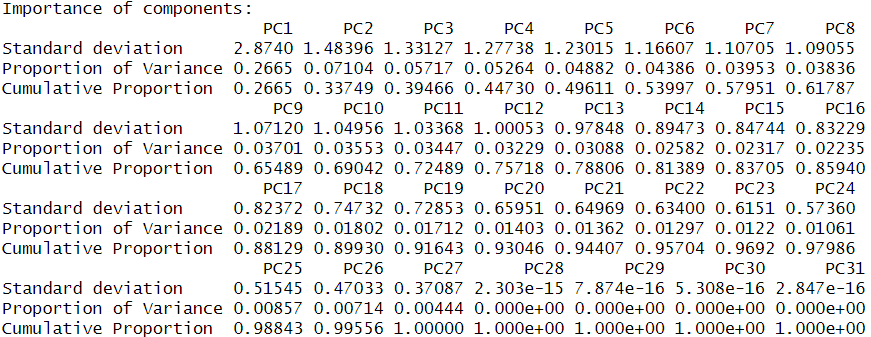
Clusters 1 and 7 were the two clusters chosen to indicate similar individuals that were likely to accept the offer, as they are both approximately 40%. Individuals within clusters 1 or 7 fulfilled the second condition that they are within clusters that have relatively high rates of acceptance. The conditions were represented as binary variables, and the two conditions were combined into one binary variable to create the list of individuals that would be targeted with future advertisement. The final list of targeted individuals was a list of 27 individuals who both were falsely predicted by the classification model to accept the offer, as well as were in clusters that had many individuals who accepted the offer. Both conditions indicated similarity to individuals who initially agreed to the offer.

**Section 4: Conclusions**

The two models outlined 27 possible individuals that could be susceptible to targeted advertising, as they are likely to make a purchase based on the criteria. This model displayed a proof of concept for combining classification models with clustering models to narrow down on individuals that are very similar to those who accepted offers from marketing efforts. Through combination of models, this process identified individuals that should be targeted with additional marketing due to their similarities to those who accepted the offer. To test whether the model was truly effective or not, it would be necessary to operationalize the results and conduct additional targeted marketing on the list of 27 individuals on the targeting list.

**Appendix**

**Figure 1: PCA variance proportions**



**Figure 2: K-Nearest Neighbors Results**

A screenshot of a computer

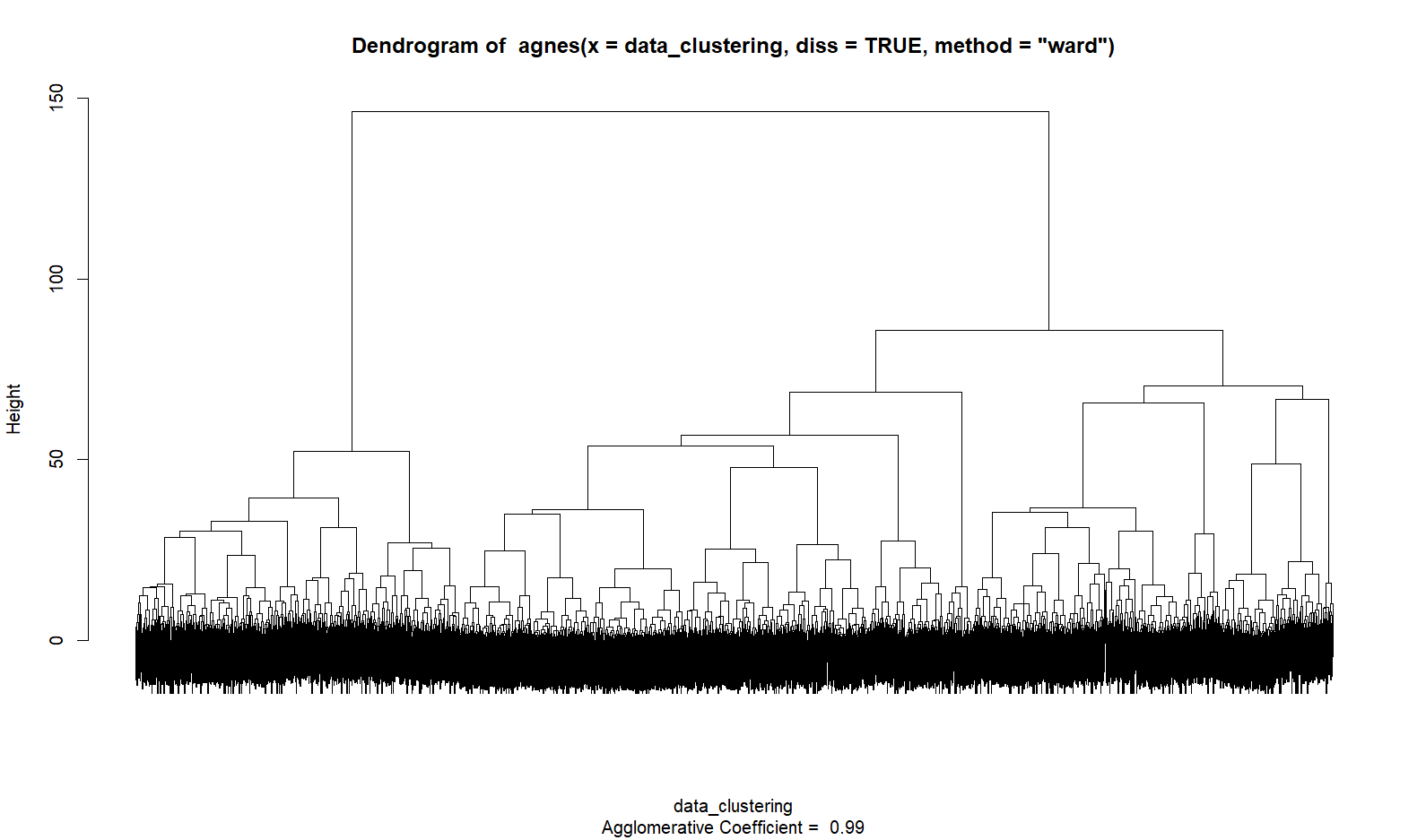
Description automatically generated

**Figure 3: Confusion Matrix for K-Nearest Neighbors Results**

A screenshot of a computer

Description automatically generated

**Figure 4: Agglomerative Clustering Dendrogram**



**References**

Daoud, Jack. “Marketing Analytics.” *Kaggle.com*, 2022,

www.kaggle.com/datasets/jackdaoud/marketing-data/data. Accessed 17 Aug.

2024.

Eser, Alexander. “Targeted Advertising Statistics Statistics: Market Data Report 2024.”

*Worldmetrics.org*, 2024, worldmetrics.org/targeted-advertising-statistics/.

Lindner, Jannik. “Targeted Advertising Statistics: Latest Data & Summary - WiFiTalents.”

*Wifitalents.com*, 5 Aug. 2024, wifitalents.com/statistic/targeted-advertising/.