“Online Retailer” Customer Segmentation and Market Basket Analysis

-Abdullah Khan, Daniel Shen, Rohan Garg, Junsu Kim

To make our clients' customer targeting strategy more efficient, we chose to analyze spending habits by region, season, and examine which items are bought together more frequently. Our group decided to move forward with a dataset of customer transactions with Online Retailer, an Australia-based E-commerce company. In this dataset, we have a list of purchases made by about 4000 customers over one year (Dec. 2010 -- Dec. 2011). The original dataset consists of 541,910 rows and 8 columns, including the invoice number, stock code, description, quantity, invoice date, unit price, customer id, and country.

There are three main objectives we wish to accomplish with this project. The first is to identify customer purchasing patterns by country. The second objective is to create customer segments to identify the most profitable customer groups to target with promotions. And lastly, we want to create and analyze product groups (market baskets) to anticipate what other products customers would buy for effective promotion campaigns.

We began by preprocessing the data. This included removing any null values/rows from the dataset as well as getting rid of duplicate rows. Additionally, we examined the summary statistics for each of the columns (predictors) in our dataset as well as box plots for the two numerical columns, “Quantity” and “UnitPrice”. After doing so, we noticed how there were about 20,000 observations of negative values for the Quantity column. This potentially could be resulting from customers returning items to Online Retailer. However, we decided to remove these observations by filtering the dataset to only include values of 1 or greater, so that they wouldn’t affect our modeling and analysis. Lastly, it is important to note that out of all the countries, the European countries seem to be the most important and relevant for Online Retailer, specifically Germany, France, Ireland, the Netherlands, and especially the UK. We can see this from the two graphs of InvoiceNo and Revenue filtered by country in the EDA process.

Seasonal and Geographical Analysis

Since we want to know more detailed information regarding the spending habits of our customers, we first merged our data to examine the effect of the time of year on sales. We observed that monthly sales peaked in November at a figure that was 20% higher than the closest sales figure from October. To get a better feel for how the season affects sales, we decided to separate the year into quarters. An interesting point we noticed was how the quarterly sales varied greatly between the first and last quarters. This might be attributed to the fact that, during winter, people are less likely to go outside to purchase goods and so they rely on an online platform instead. While this is merely conjecture, we will see later that there are some empirical results to gain insight from as well.

To take advantage of another feature in our dataset, we chose to look at demand patterns by country. Due to the sheer number of transactions in our data that took place in the UK, it comes as no surprise that sales figures are highest there. Even though we tried to control for this by log-scaling the data, the numerous purchases in the UK remain at the top.

To arrive at a concrete result, we combined our seasonal and geographic analysis. We observed that sales increased sharply in the final quarter of the year. Going from one quarter to the last, we do not see any significant changes in the order of countries' sales. However, we do see the same drastic lift in sales going into the fourth quarter of the year. So, by combining these results with information gleaned from clustering we conclude that the significant rise in sales during the final quarter of the year is attributable to the various holidays that occur at the end of the year

Creating Market Segments

Our team created customer segments using k-means clustering and then identified which goods are most commonly bought together. We used the elbow method to find the optimal number of clusters, and the silhouette method to validate these clusters and ensure the data points belong in the same segment. The silhouette method tries to maximize the similarity between data points in each cluster based on euclidean distance. This led us to six clusters of varying sizes, with almost all of them one standard deviation apart.

Among the clusters we identified, we saw varying levels of interpretability. Some clusters were very easy to group into categories such as cluster 4, which we labeled the “Celebratory Cluster”. This segment included frequently purchased items such as birthday, cake, Christmas, card, etc. Other clusters, unfortunately, were not as interpretable. An example of this is cluster 1 which we labeled bags (for lack of a better word). This segment’s most frequent terms included vintage, paisley, bag, and storage. While these items may not tell us much through their relation to each other, the fact that those items landed in the same cluster allows us to infer that buying these items together is so popular among customers that it may be an ideal opportunity to target these customers with a joint promotion. Another very interesting segment was cluster 5, labeled “Home, Garden, & Tools” which saw the pairing of home goods with gardening goods, another potential avenue for a joint promotion between product categories.

Market Basket Analysis

After generating product categories through K-means clustering, we can optimize OnlineRetailer’s customer retention and profitability by conducting market basket analysis and recommendation systems. Before constructing the model, we first grouped our dataset by invoice and performed one-hot encoding to get dummy variables indicating whether the invoice contained specific product categories or not. Having the dataset processed, we first computed the purchase occurrences by summing the individual purchase incidents. Then, we calculated the purchase co-occurrences by

summing up all the co-purchase incidents. Next, we obtained the S_ab score by dividing the co-occurrences by the square root of the product of occurrences of the two categories. Finally, we identify the corresponding S_ab score based on categories the customers have purchased, and the category having the largest S_ab score is the product we are recommending to OnlineRetailer's customers.

Our final product is a robust recommendation system trained with 541,909 purchase incidents. By inputting the product categories the customer purchases, it outputs the recommended category out of the remaining ones.

Conclusions/Recommendations

Through customer segmentation, we created product categories based on the most frequently purchased items. If Online Retailer does not have a customer targeting strategy already, they could benefit greatly from separating their customers into the segments that we found using k-means clustering. That in conjunction with the market basket analysis provides us with a variety of actionable insights for our client. If Online Retailer does already have certain segments or product categories in place then they can cross-validate using ours.

Online Retailer will be able to better anticipate the categories customers are likely to purchase next with our proposed recommendation system. There are plenty of potential areas where Online Retailer can incorporate the anticipated customer purchase. For example, implementing the promotion of recommended products and having a recommended product section when customers are checking out on the website are both opportunities to utilize these insights. Furthermore, we would recommend that Online Retailer tailor their marketing strategy based on customer demand by season, as this will allow them to capitalize on their increased demand during the fourth quarter of the year. By combining all the findings in our analysis, Online Retailer is left with a robust targeting strategy at their disposal that could lead to strengthened customer loyalty, and (in most cases) an increase in profitability.