**Course Project Final Report**

Group 3

Shawn Anderson, Greg Funaro, Lorenzo Gordon, Timothy OMeara & Sarah Vinyard

**Project Overview and Dataset**

Do certain country buying patterns vary against others? To support our initial question, we are also asking: Do buyers from certain nations buy more of one product over another, or none at all? Do they buy during certain times of the year over others, or are their buying patterns consistent? This project aims to investigate a dataset that contains certain attributes of a wholesale buyer, their country and the products they purchased and will use various data mining techniques to answer this initial question.

For our course project we chose an [online retail data set](https://archive.ics.uci.edu/ml/machine-learning-databases/00502/online_retail_II.xlsx) from the UCI repository. It has eight attributes and 541,909 instances in the dataset. This is a transactional dataset which contains all the transactions occurring between 1/12/2010 and 9/12/2011 for a UK-based online (no brick and mortar) retail store. The company mainly sells unique all-occasion gifts and many of the customers are wholesalers.

The eight attributes are outlined below:

* InvoiceNo *(nominal)*: 6-digit number uniquely assigned to each transaction
* StockCode *(nominal)*: 5-digit number uniquely assigned to each distinct product
* Description *(nominal)*: product (item) name
* Quantity *(numeric)*: quantities of each product (item) per transaction
* InvoiceDate *(numeric)*: day and time when each transaction was generated
* UnitPrice *(numeric)*: product (item) price per unit in sterling
* CustomerID *(nominal)*: 5-digit number uniquely assigned to each customer
* Country (nominal): name of the country where each customer resides

**Preprocessing**

Before we were able to import our dataset into Weka and RStudio, we first had to clean our variables to replace/omit special characters (ie. , = “ ‘ \* + - %).

We cleaned out any entries that had a blank/na/Unspecified CustomerID or Country as this would skew the data we wanted to use to answer our business question.

Additionally, we added the column ‘Continent’ (for which there were 6) to yield more balanced results. Before the additional column, it was hard to identify where the trends were and where to place the centroids. In order to match countries to each of these continents, we changed the following “countries” to be their country proper:

* Hong Kong → China
* Channel Islands → United Kingdom
* EIRE → Ireland

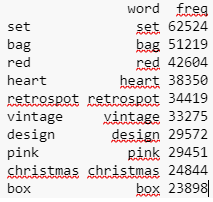
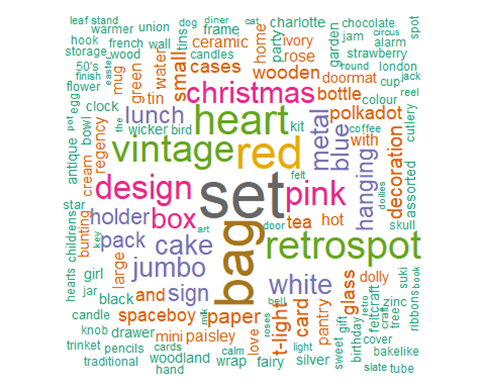
To make analysis easier, we converted InvoiceNo, StockCode and CustomerID from string to nominal and Quantity from numeric to nominal. These changes will aid us in our data mining tasks.

**Text Mining**

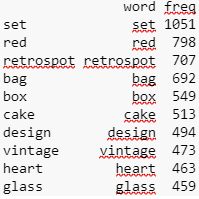
Before moving on to our prediction models, we decided to convert the descriptions from the .csv into .txt files in order to gain a better understanding of the kind of data we were working with, and see early on if there were any particular product descriptions that stood out overall. We hoped that this would help us in our analysis of the models that follow.

We split the descriptions into sales and returns; based on the quantities (interpreting any quantities with a negative number to be a return).

*Sales:*



*Returns:*



It was discovered that there was significant overlay in the words used for descriptions for both sales and returns. To identify where this overlap differed, we looked to find the association between some of the top words.

*Associations*

|  |  |  |
| --- | --- | --- |
|  | **Sales** | **Returns** |
| *Set* | N/A | Tins - .42 |
| *Bag* | Jumbo - .55, Lunch - .41, Charlotte - .31 | Jumbo - .56, Lunch - .46, Charlotte - .34 |
| *Red* | Retrospot - .48 | Retrospot - .54 |
| *Christmas* | 50s - .39 | Lights - .36, 50’s - .32 |

We found that words like “Set” didn’t have any strong associations with other words when it came to Sales. However, it is associated with words like “tins” regarding

Returns. Words like “Red” and “Retrospot” tended to mainly be associated with themselves. Similarly, “Bag”, for both Sales and Returns, was associated with words like jumbo, lunch, and charlotte. The first instance of differences with any significance was with the term “Christmas”, where we saw that Returns tended to associate with both “50’s” and “Lights” whereas Sales only associated with “50’s”, which can lead to the assumption that people often return Christmas lights.

When splitting the descriptions based on their countries, there was still minimal variation from this. For instance, the UK (making up 496,237 of the observations within the data set) maintained the same high frequency words that countries such as Ireland did with 8197 observations.

*UK:*

**Word freq.**

set 56343

bag 47734

red 38415

heart 37146

Vintage 31069

*Ireland:*

**Word freq.**

set 1087

red 705

vintage 578

retrospot 533

heart 476

Based on what we’ve looked at so far, considering the terms that appeared most frequently, as well as the associations between them, it does not appear that there are many strong connections between descriptions and the likelihood of sales, returns, or patterns by country.

**Classification**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Accuracy** | **Parameters** |
| *Association Rule Mining* | See Selected Rules | See Plot1 & Plot2  Min Support = 0.01  Min Confidence = 0.10  Min Length = 3-6  Max Length = 4-8 |
| *SimpleKMeans* | 43.8% | numClusters: 6; seed=10,000;  Classes to clusters evaluation |
| *J48 Decision Tree* | 99.7% Country Prediction  72.67% for Customer Frequency  45.54% quarter | C-0.25, M-2  C-0.5 m-12  C-.5, m-12 |
| *SVM* | 33.5% | \*\*5,000 sample size\*  kernal=polykernal;  exponent=2.0;  c=1  trainingPercent=80%;  Run Time = 252 seconds |

**Association Rule Mining (ARM)**

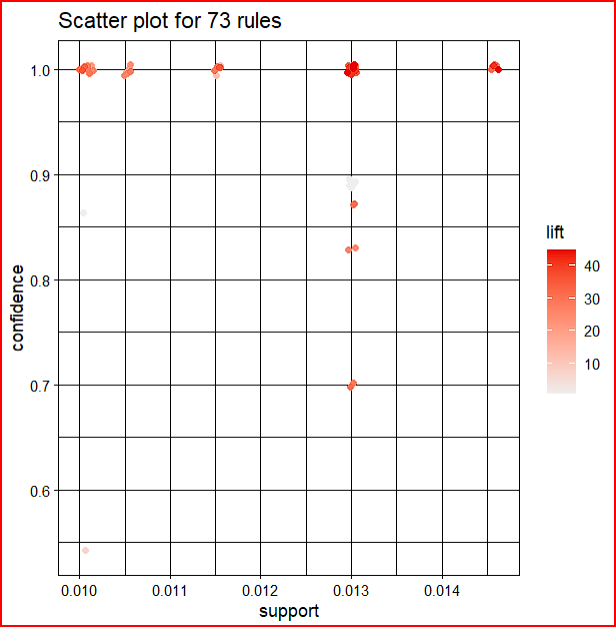
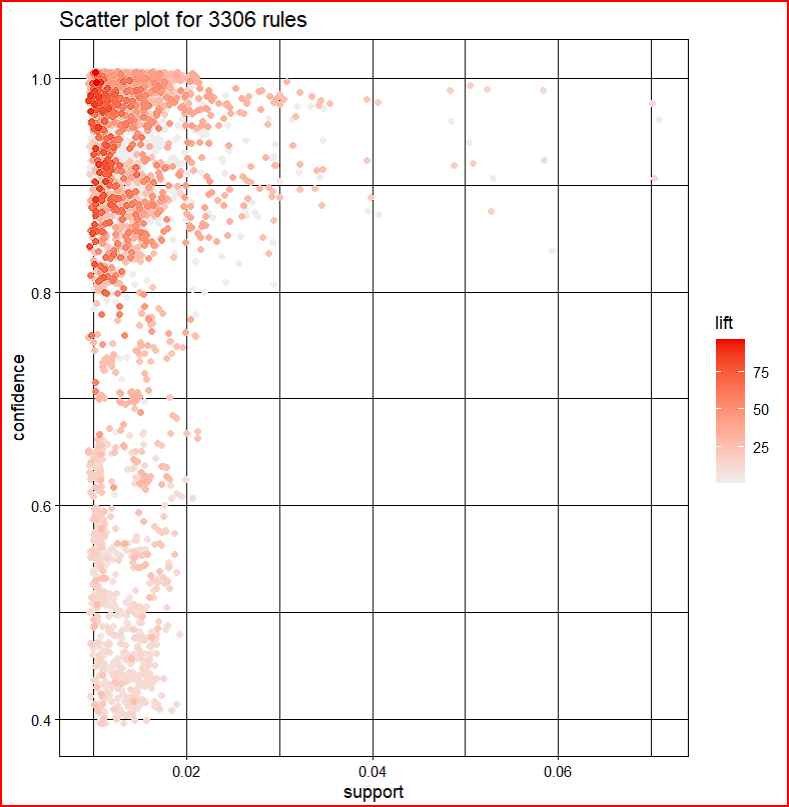
*NOTE: StockCode and Description are the highest information gain elements and were both retained to more easily interpret rules. Additionally, a unique StockCode does not guarantee a particular Description in all cases due to many adjusting accounting entries and seemingly insufficient data validation practices.*

We first evaluated the general rule environment with filtering parameters (parameters listed in table above with length of 3:4 elements). We then created a 3-dimensional plot of this in order to assess the relative strength and interestingness of over 3,000 rules (Plot1). This allows us to strategically work with management in setting refined parameters to visually catch the more interesting rules - and bypassing those which first seem less promising/actionable. Rules in the upper-most left region and along the 1.0 Confidence axis are therefore ignored for this report due to their high correlation with paired StockID/Descriptions and related CustomerID/Country features.

Plot2 was therefore grown to 6:8 items (73 rules) to leverage greater information gains through the potential of additional Stock/Description elements such as:

minlen = 6: CustomerID, Country. StockCode1/Description1; StockCode2/Description2

maxlen = 8: e.g. creates room for an additional StockCode3, Description3

Plot1 (3:4 Rule Length) Plot2 (6:8 Rule Length) 

This change allows Plot2 to highlight a vertical pocket of rules along the 0.013 Support vertice which suggest a ~24-30x lift for a third type of teacup and saucer set after UK buyers add two other unique sets.

*RECOMMENDATION: Management may consider acting on this information by prominently displaying special offers for high-margin teacup products or other closely associated high-profit items in the shopping window as customer carts meet these purchasing patterns.*

{22697, 22699, GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER, United Kingdom} => {22698 “PINK REGENCY TEACUP AND SAUCER} Support: 0.01 - Confidence 0.70 - Lift: 31.3

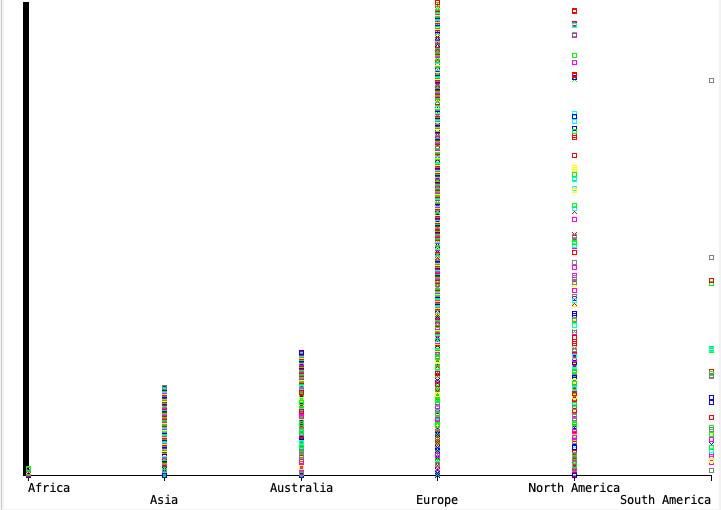
**Clustering**

*Centroids K-Means Clustering with clusters set to 5 and 66% split*

* Cluster 0: Europe, makes up 99.4% of purchases, top item purchased: “white hanging heart t light holder”, biggest purchasing time of year: Q4 (34%)
* Cluster 1: Africa, makes up 0.01% of the purchases, top item purchased: “spaceboy wall art”, biggest purchasing time of year: Q4 (100%)
* Cluster 2: North America, makes up 0.08% of the purchases, top item purchased: “tea party birthday card”, biggest purchasing time of year: Q4 (61%)
* Cluster 3: Australia, makes up 0.23% of the purchases, top item purchased: “set of 3 cake tins pantry design”, biggest purchasing time of year: Q3 (35%)
* Cluster 4: Asia, makes up 0.24% of the purchases, top item purchased: “manual”, biggest purchasing time of year: Q3 (35%)
* Cluster 5: South America, makes up 0.01% of the purchases, top item purchased: “set/3 red gingham rose storage box”, biggest purchasing time of year: Q2 (100%)

*Nation Cluster Breakout*

* Europe: Austria, Belgium, Cyprus, Czech Republic, Denmark, European Community, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Lithuania, Malta, Netherlands, Norway, Poland, Portugal, Spain, Sweden, Switzerland, United Kingdom
* Africa: RSA
* North America: Canada, USA
* Australia: Australia
* Asia: Bahrain, China, Israel, Japan, Lebanon, Saudi Arabia, Singapore, UAE
* South America: Brazil



We chose to do 6 clusters for SimpleKMeans, one for each continent, instead of letting the algorithm determine the number of clusters. We did this mainly because of the number of countries, 37, and the limit of compute power. This also helped bulk up some of the smaller clusters from being so heavily skewed in one direction.

Cluster 0 is defined as having purchases predominantly from Europe, and the distinguishing feature for this cluster is the count of Invoices. Europe is listed as having the most invoices (largely due to the United Kingdom, Germany, France and Ireland), the largest number of Stock Items per invoice and the largest dollar amount spent per invoice. The top item that Europe is purchasing is the “white hanging heart t light holder”. In addition, Europe is fairly consistent in their purchasing time of year. However, Q4 has a slightly larger purchasing time period (made up of 34% of yearly sales).

Cluster 1 is one of the smallest clusters within the set, and is defined by the small amount of invoices purchased by the continent. Africa makes up 0.01% of the total invoices purchased in 2011 and only contains one country (RSA) which correlates with the limited number of invoices. The top item that Africa is purchasing is the “spaceboy wall art”. Africa made all 58 of their purchases in Q4.

Cluster 2 only contains two countries (USA and Canada), but has the third largest number of purchases of all of the continent clusters. Their top item purchased was the “tea party birthday card” and also saw a spike for their purchases around Q4 (60.8%), but Q2 and Q3 were tied for second.

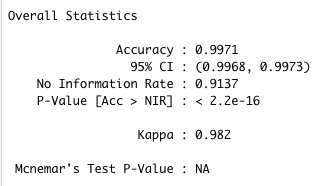
Australia makes up cluster 3 alone, and has the second largest count of invoices (making up 0.23% of total invoices). Each invoice on average contains 183 items and the top item that Australia is purchasing is the “set of 3 cake tins pantry design”. While Q3 has the highest number of invoice purchases (35%), Q1 and Q2 follow closely in pursuit.

We define Cluster 4 as the Asia continent, and having the second largest number of countries contributing to the number of invoices created. Australia and Asia have similar buying patterns (ie. average quantity of invoices items, average invoice amount, purchasing periods, etc.). This could likely be because of their proximity in location, but we would need to do more research before saying this for certain.

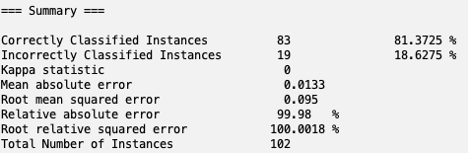
Cluster of 5, South America, is mainly defined by its lack of attributes as it only has 32 invoices (the smallest number of invoices of any cluster). Additionally 24 items were every purchased, making the ratio of repeat purchases basically zero (ie. it almost seems like they are testing out items for purchasing). The top item that South America is purchasing is the “set/3 red gingham rose storage box”, and they only purchased items in Q2.

**Decision Tree Theory**

For our decision tree model the first task was to see how well this model can predict the country that these orders originated from. We initially decided to use the default J48 setting as a baseline for the prediction. With the default rweka controls the training data was able to predict 99.7% of the training data.

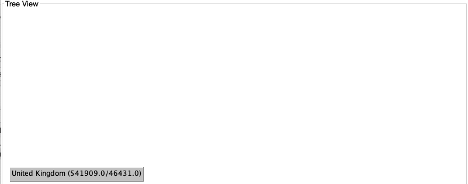


Next we looked a little further into the data and info gain that each attribute generated towards the country the top variables were, in InvoiceNo(.74), InvoiceDate(.72), and customerID(.71). This makes intuitive sense as the customerID and location are things that may remain constant, so if a customer typically orders from China it would be safe to assume that they will continue to order from there. We wanted to see if removing these would keep the same accuracy. So after removing those 3 variables we ran the dataset through a 3 fold cross validation where the prediction resulted in 91% accuracy, however the size of the tree was only 1. Tested this model against a dataset from the same retailer, but from the previous year which resulted in a 81% prediction accuracy.



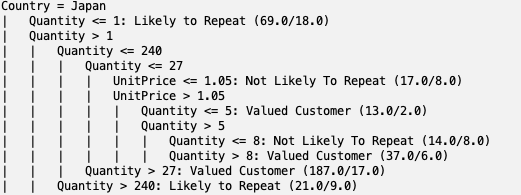
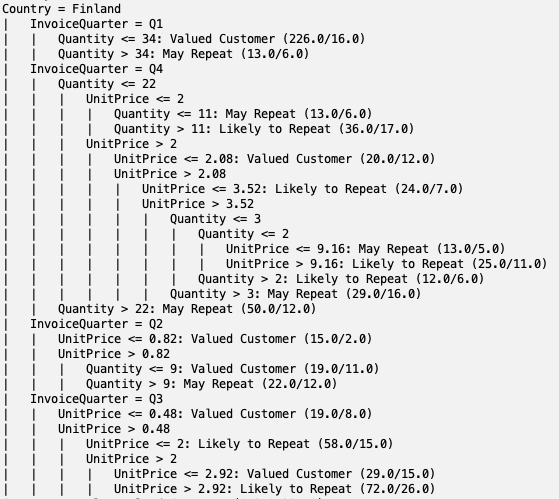
*(Prediction against previous year)*

Looking further into the data a bulk of our countries are the United Kingdom(91%) and when looking at the tree that the decision tree created it was only the United Kingdom. Our decision tree was only picking the united kingdom as a country and because the dataset is heavily weighted by the united kingdom values it led to a 90%+ accuracy. Therefore determining the country based on the other variables may not be the best route with this training data.



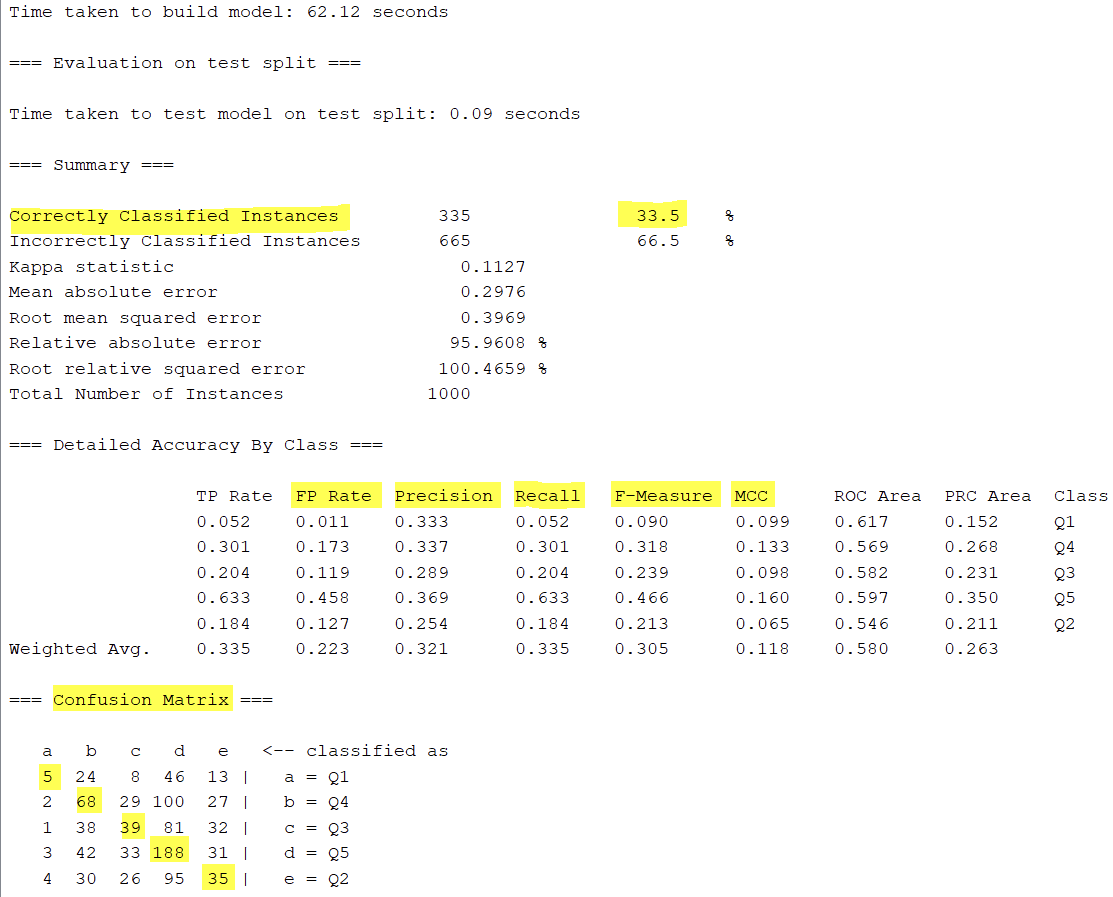
*(Decision tree output)*

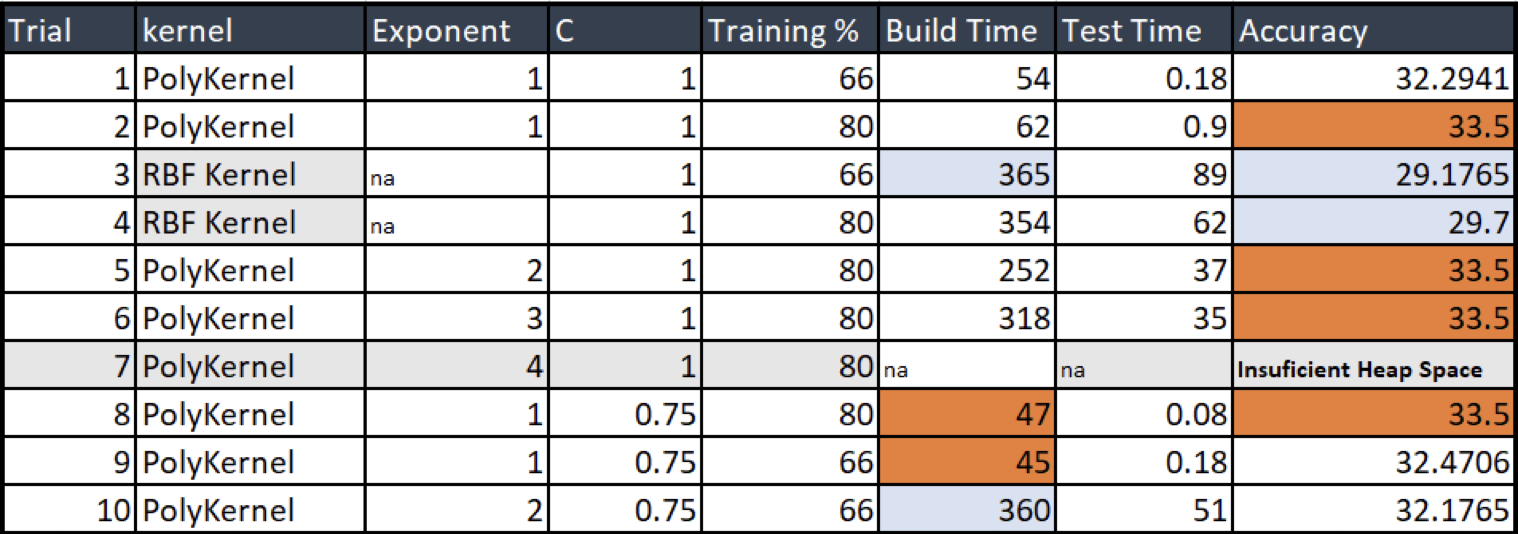
Moving from the picking country we investigated Customer frequency and invoice quarters. Customer frequency was determined combing the # of times a customer’s ID appeared then finding the range of the customer which goes from 1 purchase to 7772 purchases. Next we found the quartiles for the customer frequency and assigned a title for each to represent their value. The groupings were “Not Likely to Repeat”, “May Repeat”, “likely to repeat”, “Valued customer.” The InvoiceQuarter column was created by extracting the month from the invoicedate and creating bins for quarters in R. invoiceDate was also removed before starting the analysis on rweka. Using this new information we attempted to predict Customer Frequency with a 3 Cross validation and we received a prediction accuracy of %72.5. After adjusting the confidence factor to 0.5 and minobjects to 12 the accuracy moved to 72.67%. Country is still a key variable. However, we are able to see further breakdowns by country and quarter. An example of this can be seen with countries like Finland and Japan



*(Decision Tree for Customer Frequency)*

**SVM**





**Conclusion**

There are a vast array of buying patterns between continents with Europe having the most activity. Invoice overlap between cluster 3 and 4 is likely due to location proximity.

*RECOMMENDATIONS*: Christmas lights were found to be a major return item after the holidays and might therefore warrant modified return policies. Additionally, management may wish to consider adjusting the online shopping experience as suggested in the ARM section to upsell higher profit teacup and saucer products to their UK customers.