Categorizing Customer Value and Exploratory Analysis: Semi-Supervised Machine Learning Applied to the UCI Machine Learning Repository Online Retail Dataset.

Miley H. Thompson

University of Oklahoma School of Library and Information Studies

Introduction

Background

The University of California Irvine (UCI) Machine Learning Repository has published a dataset of transaction records, from 01/12/2010 to 09/12/2011, of a UK based business (Online Retail). This data contains useful information including products purchased, the value of the products, the quantity purchased, the date of the transaction, and the country the purchaser resides in. The data set may be analyzed from many angles to create insights the business should use to optimize their operations. Some potentially relevant analyses include: product purchase frequency in relation to date, customer, country, and product price; product purchase frequency in relation to other products i.e. which products are often purchased together; and customer purchasing habits, and their overall value brought to the business. These will require a range of techniques such as descriptive statistics, feature engineering and machine learning modeling.

Purpose

In this project, I will perform surface level analysis of the dataset, which will provide immediately applicable—but limited in precision—insights, and an in-depth analysis of the customers purchasing behavior. The in-depth analysis will provide groupings of customers based on purchasing habits, and a model to classify the customer's into low, moderate, and high value purchasers. This model should be used to inform the business in multiple ways including sales efforts, and the prioritization of operation efforts. Finally, this project, complete with an extensive understanding, organization, and preparation of the data, will provide a groundwork for efficient future analyses to answer other questions and further optimize the business' operations.

Data Understanding

Before work may be initiated, a comprehensive understanding of the dataset must be gained. First, I will break down general information about each feature of the data. Then, I will explain what issues exist in the data's current form.

Data Features

1. Invoice number. This represents a specific occurrence of an order. It is shared between the observations that make up a single order.
2. StockCode. This represents a unique identifier for a single product. It is in the form of a 5-digit number plus a single uppercase character. The character represents different versions of a similar product.
3. Description. This represents a short description or name of a product.
4. Quantity. This represents the quantity of the item involved in the transaction. This can be both a positive or negative number.
5. InvoiceDate. This represents the date at which the transaction occurred. It is in the form: "mm/dd/yyyy hh:mm"
6. UnitPrice. This represents the purchase price of the item purchased in the observation. It is measured in the Pound Sterling.
7. CustomerID. This represents the individual customer that has made the purchase.
8. Country. This represents the country in which the purchaser resides.

Essential to understanding the data is considering its shape, that is the form it takes and the type of information it conveys. Initially, the data is shaped that each observation represents a change of inventory of some quantity of a product. The term transaction is used to represent this. However, it is slightly deceptive because it incorrectly infers that a purchase has been made. Rather, a 'transaction' can represent a purchase of a product, a return of a product, or the unsold loss of a product. Further, while the data

currently is shaped that it shows transactions, contained in the data are structures that might inform very different ideas. For instance, the data can be reshaped so that each observation is a specific product, or—as I will perform—each observation represents a specific customer.

Problems in the Data

While the integrity of data is generally high, with only the CustomerID feature containing a notable number of missing values, there remains a few issues that should be resolved before the data is subset.

1. The InvoiceDate feature contains more information than necessary for this analysis.
2. The features Description and StockCode are intended to be directly correlated; every unique Description value should connect to a single StockCode value and vice versa. Yet, 128 Description values have multiple StockCode values. Of these 128 instances, 1 is due to an empty description value, 104 are due to data entry errors in the capitalization of the value, and the remaining 23 are actual separate StockCode values.
3. The CustomerID feature contains 135,080 observations with missing values.

Further, as this project is specifically concerned with the actual purchases made and the customers behind those purchases, the data will need to be subset to remove any non-purchase observations, and reshaped so that it represents the customers rather than transactions.

Methods

This project uses two tools to prepare its data and perform analysis. Excel is used for quick, visual data exploration and the R scripting language is used for the bulk of transformations, modeling, and visualizations. Within the R scripting, the following packages are essential for this project: tidyverse, reshape2, ggplot, ggally, treemapify, rpart, caret, imbalance, metrics, and vip.

Data Preparation

Data Cleaning 1

Two changes are applied at this step. First, as the scope of this project's analysis limits the acceptable level of detail, excel functions are used to transform all InvoiceDate values into Month-Year format. Second, to begin resolving correlation errors between StockCode and Description, all character values in StockCode are transformed into uppercase characters.

Data Subsetting

As previously described, the data must be subset to include only observations representing actual purchases. These observations are easily distinguished by having whole number, positive Quantity values and a UnitPrice value above zero. The following formulas are used to split the data:

a. purchases = all observations where Quantity >= 1 and UnitPrice > 0
b. non_purchases = observations where Quantity < 1 or UnitPrice <= 0

This results in two subsets: Purchases containing 530,104 observations, and non-purchases containing 11,805 observations. Next, the missing CustomerID values must be resolved. There is not a clear reason as to why there are missing values and there are no patterns in these observations that reveal a reasonable imputation strategy. Therefore, all observations missing a CustomerID value are removed. This results in a subset of 397,884 observations.

Data Cleaning 2

Finally, after subsetting, there remains 17 Description values that relate to multiple StockCode values. Without a revaluation of the data collection process—which is far outside the scope of this project—it is

impossible to infer why these errors are present. So, to maintain direct correlation between StockCode values and Description values, for every error, the most common StockCode value, for a given Description value, will be imputed onto the other StockCode values for that Description value.

For example, if Product A has 2 observations with a StockCode value of 123 and a third observation with the stockcode value of 124, the StockCode value of the third observation will be replaced with the value 123.

Data Reshaping and Feature Engineering

Now that the data has been cleaned and subset, the next stage of preparation is to reshape the data. As my goal is to examine the customer habits, I will reorient the data so that every observation represents information on a single customer. The first step of this is to create a new feature in the purchases subset which represents the monetary value of each transaction. This is done by taking the product of the Quantity value and the UnitPrice value for each observation. Next, I will create a new dataset populated with each customerID value and the number of transactions associated with that value. Subsequently, new features are added to this table. The resulting dataset summarized below.

1. CustomerID. This represents a single customer. Every observation contains a unique CustomerID.
2. NumTransactions. This represents the total number of records associated with each CustomerID.
3. AvgUnitPrice. This represents the average price of the items purchased by the customer.
4. NumOrders. This represents the number of orders the customer has placed.
5. AvgOrderSize. This represents the average number of items in that customer's orders.
6. SumOrderValue. This represents the total value of all goods purchased by the customer.

K-means Model Preparation

Next, for modeling purposes the data in customer data must be normalized. The scale function is applied so that every feature has a mean of 0 and a standard deviation of 1. This will allow the effective application of the k-means clustering algorithm.

Data Modeling

This project takes a two-step process to creating the customer classification model; two styles of machine learning will be used and combined to create a nuanced model specific to this dataset. First, the k-means clustering algorithm will be used to separate the customer data into three groups. These groups themselves will be analyzed, then they will be used as labels for a Classification And Regression Tree (CART) algorithm to produce a model that will reveal impactful features, and efficiently categorize new customer data in the future.

K-means Clustering

To create the k-means clustering model, first the k value, which is the number of groups, must be determined. For this project it is intuitive to set the group number at 3 as that is the number of customer categories we are looking for. Further, testing various k values reveals that having a larger k value creates groups that are too small and with differences to minute for analysis. The k-means function of the R cluster package is used to create the groups. The differences in each group will be explored in the results section.

CART Model Preparation

Now that the data is labeled into three groups, referred to as classes, it must be split into a training set and a testing set. This is done via stratified, random partitioning using the caret package. Stratified means that the proportion of each label is maintained in each partition. The training set contains a randomly selected seventy percent of observations while the testing set contains the remaining thirty percent of observations.

However, I will be using a CART based decision tree model. This type of machine learning is sensitive to imbalanced data. Data imbalance refers to having inequal amounts of each class in the data. Currently, the data is heavily imbalanced. Within the training set there are 2,487 instances of class one, 541 instances of class two, and 9 instances of class three. The strong minority of class three will result in it being disregarded by the model—this must be avoided. To correct this, I will perform Random Walk Oversampling (RWO). Oversampling is a technique in which the minority class is either counted multiple times or used to synthesize similar data. Random Walk refers to the specific type of oversampling. This style was chosen because RWO emphasizes maintaining the same mean and deviation values in the resulting data. Because I am interested in the summarized data, maintaining these values is crucial.

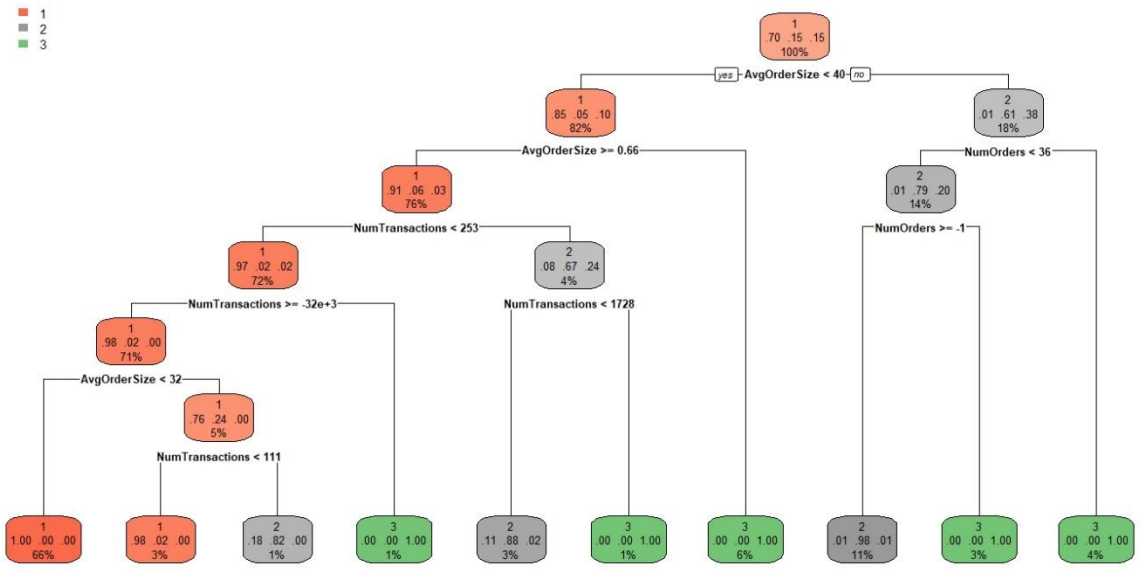
RWO is applied to the minority class so that the training data contains 2,487 instances of class one, 541 instances of class two, and 541 instances of class three. Note, the data is still imbalanced. However, it is balanced enough that the model will consider each class as relevant, which it previously would not, and the overall imbalance is still representative of the actual data. Because the classes were assigned based on the actual groupings of the data, to completely balance the training data would be to introduce an unreasonable amount unrealistic information—data that is not representative of real life.

Decision Tree Modeling and Testing

Now that the data is prepared and partitioned, the CART decision tree model is created using the rpart package in R. This results in a decision tree that sorts the customers into their classes. This decision tree is shown in figure 1.

Figure 1

CART Decision tree model for customer classification.



To establish the validity of this model, the testing data will be fed into the decision tree and predict the classifications of the customers. Note that while the model was trained on oversampled, relatively balanced data, the testing data is unbalanced but stratified—containing a ratio of classes representative of the true data. Therefore, the predictions are reasonably accurate to a true test with new, unlabeled data, if we assume that the new data is of a similar nature to the current data.

To assess the model, the predicted values are compared to the true values in a confusion matrix. The confusion matrix tells us how many true and false positives and negatives are present in the predictions. Table

To assess the model, the predicted values are compared to the true values in a confusion matrix (Table 1). The confusion matrix tells us how many true and false positives and negatives are present in the predictions. Table shows this confusion matrix along with the notable model testing metrics of accuracy and Root Mean Square Error(RMSE).

Table 1

Confusion Matrix for CART Model

Confusion Matrix	true values				
predictions	1	2	3		Accuracy
1	1030	6	1		0.9785
2	18	242	3		RSME
3	0	0	1		0.154363

Notice the especially high accuracy score and the especially low RSME. These metrics demonstrate that the high validity of the CART model

Results and Analysis

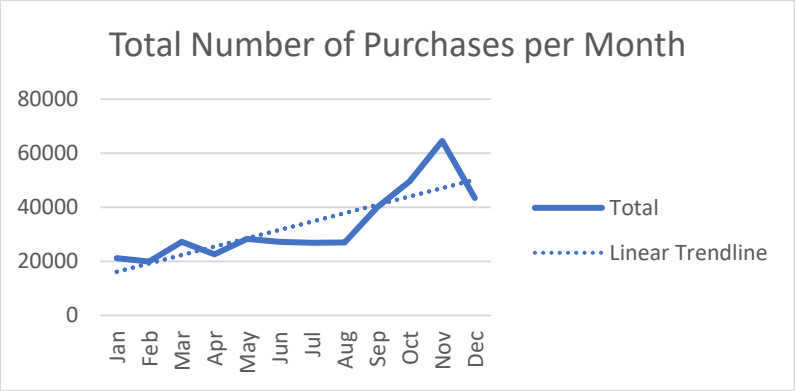
Throughout this project, three different perspectives have been revealed. First, the frequency of purchases and frequency of specific product purchases tell us what products are most demanded and when they will be demanded. Second, the locations where purchases come from inform us what geographic markets the business is a part of and where they should expand to first. Third, the purchasing habits of individual customers show what kinds of behaviors can be expected from different types of customers and which customer should be prioritized in marketing and operational efforts.

Frequency of Purchases and Products Purchased

First, I will consider when the business receives most of its orders and explore how this might inform operations decisions. Figure 2 shows the total number of purchases occurring each month, and the overall trend of purchases.

Figure 2

Plot of Sum Purchases by Month



This suggests that the expected number of orders remains relatively consistent, from the months February to August, then between August and November sales sharply increase, before sharply decreasing in December and January. I assert that this trend is due to sales increasing for the December holiday season. It is initially surprising that the descent begins in December rather than after December—when consumers will have presumably finished holiday shopping. However, the majority of the businesses purchases are from wholesale customers—not individual consumers (Online Retail). I suggest that this phenomenon is due to wholesale customers stocking up in preparation for the height of seasonal sales—hence the early increase in purchases.

Now, I consider which individual products are most commonly purchased. Table 2 shows ratio and count of the 10 most frequently purchased products. Figure 3 further visualizes these products ratios.

Table 2

The top 10 most frequent Items (Description) purchases and their ratio to the total number of purchases.

Description	ratio	count
WHITE HANGING HEART T-LIGHT HOLDER	0.51%	2028
REGENCY CAKESTAND 3 TIER	0.43%	1723
JUMBO BAG RED RETROSPOT	0.41%	1618
ASSORTED COLOUR BIRD ORNAMENT	0.35%	1408
PARTY BUNTING	0.35%	1396
LUNCH BAG RED RETROSPOT	0.33%	1316
SET OF 3 CAKE TINS PANTRY DESIGN	0.29%	1159
LUNCH BAG BLACK SKULL.	0.28%	1105
POSTAGE	0.28%	1099
PACK OF 72 RETROSPOT CAKE CASES	0.27%	1068

Figure 3

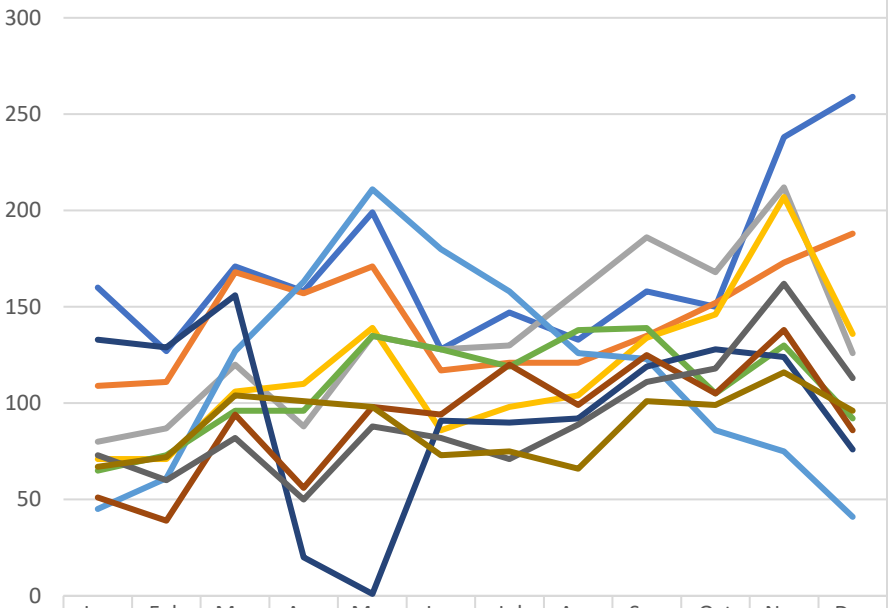
Treemap visualization of the 10 most frequently purchased items.

While I choose to only report the top 10 products, an extensive list is easily producible and may be used to inform which products are should be kept in stock.

Lastly, these two approaches may be combined to show nuances in the effect of time on product purchase frequency. Figure 4 shows the purchase amounts of the top 10 products based on month.

Figure 4

Plot and Table of the Top Ten Most Purchased Products Frequency by Month



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
WHITE HANGING HEART T-LIGHT HOLDER	160	127	171	158	199	128	147	133	158	150	238	259
REGENCY CAKESTAND 3 TIER	109	111	168	157	171	117	121	121	135	152	173	188
JUMBO BAG RED RETROSPOT	80	87	120	88	135	128	130	158	186	168	212	126
ASSORTED COLOUR BIRD ORNAMENT	71	71	106	110	139	86	98	104	134	146	207	136
PARTY BUNTING	45	61	127	163	211	180	158	126	123	86	75	41
LUNCH BAG RED RETROSPOT	65	73	96	96	135	128	119	138	139	105	130	92
SET OF 3 CAKE TINS PANTRY DESIGN	133	129	156	20	1	91	90	92	119	128	124	76
LUNCH BAG BLACK SKULL.	51	39	94	56	98	94	120	99	125	105	138	86
POSTAGE	73	60	82	50	88	82	71	89	111	118	162	113
PACK OF 72 RETROSPOT CAKE CASES	67	72	104	101	98	73	75	66	101	99	116	96

While the overall trend shown in figure 2 is somewhat present, there are specific products, such as “SET OF 3 CAKE TINS PANTRY DESIGN” and “WHITE HANGING HEART T-LIGHT HOLDER” that are exceptions to this trend. This suggests that multiple trends may exist and a larger, product-oriented analysis should be done. This will be further explored in the discussion.

Locations of Customers and Market Implications

Next, the location of purchasers is explored. Being a United Kingdom (UK) based business, it may be assumed that the majority of orders they receive are from UK customers. The data confirms this. The vast majority, 89%, of transactions occur with UK based customers. Table 3 shows the ratio and count of

transactions based on country. Figure 5 visualizes the locations of these countries and their relative ratio of transactions.

Table 3

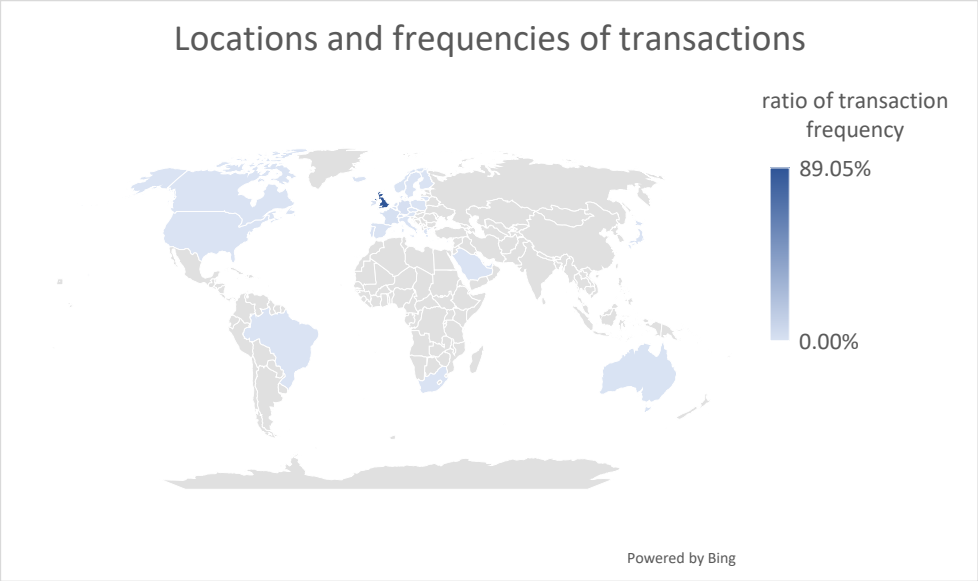
Ratio of transaction frequency and Sum of Transactions by Country

country	ratio of transaction frequency	count
United Kingdom	89.05%	354321
Germany	2.27%	9040
France	2.10%	8341
EIRE	1.82%	7236
Spain	0.62%	2484
Netherlands	0.59%	2359
Belgium	0.51%	2031
Switzerland	0.46%	1841
Portugal	0.37%	1462
Australia	0.30%	1182
Norway	0.27%	1071
Italy	0.19%	758
Channel Islands	0.19%	748
Finland	0.17%	685
Cyprus	0.15%	614
Sweden	0.11%	451
Austria	0.10%	398
Denmark	0.10%	380
Poland	0.08%	330
Japan	0.08%	321
Israel	0.06%	248
Unspecified	0.06%	244
Singapore	0.06%	222
Iceland	0.05%	182
USA	0.04%	179
Canada	0.04%	151
Greece	0.04%	145
Malta	0.03%	112
United Arab Emirates	0.02%	68
European Community	0.02%	60
RSA	0.01%	57
Lebanon	0.01%	45
Lithuania	0.01%	35

Brazil	0.01%	32
Czech Republic	0.01%	25
Bahrain	0.00%	17
Saudi Arabia	0.00%	9

Figure 5

Geographical Heat Map of Transaction Frequency by Country



With this information, I infer two perspectives.

1. To maintain the businesses current value, that is to avoid losing value, resources should be prioritized toward the UK customers.
2. Currently, the international market is not being properly tapped by the business. To grow the businesses value, marketing efforts should be created to increase purchases in the international market. As a first step toward this, the international countries with the highest purchasing ratio should be prioritized first. Table 2 shows these to be Germany, France, and Ireland (EIRE), each providing at least 1.8 % of total transactions.

Customer Purchasing Habits

Finally, the bulk of this project is the customer purchasing habits and classification of their value to the business.

In its original state, meaning unclustered, the data has a very high variability. Table 4 shows summary statistics of each feature. Figure 6 shows boxplots, representing these summary statistics of each feature.

Table 4

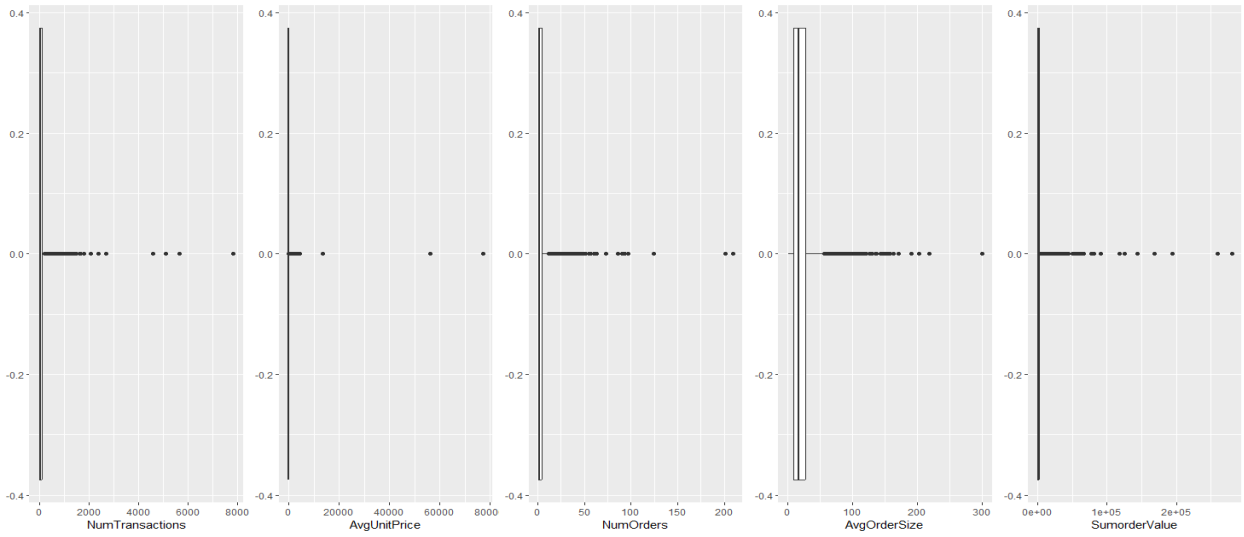
Summary Statistics of Customer Purchasing Habits

All Customers	NumTransactions	AvgUnitPrice	NumOrders	AvgOrderSize	SumorderValue
min	1	2.101285714	1	1	3.75

1st quarter	17	12.3653667	1	9.5	307.415
median	41	17.72311929	2	17	674.485
mean	91.72060858	68.3505056	4.272014753	22.19017297	2054.26646
3rd quarter	100	24.85841667	5	28.25	1661.74
max	7847	77183.6	209	300.6470588	280206.02

Figure 6

Boxplot of Customer Purchasing Habits



Notice, there are very strong outliers in the data. These represent valid data points but obscure the interpretability of these statistics. This current state supports the idea that the data should be grouped together before analysis. k-means clustering results in our desired three groups. Table 5 presents these clusters.

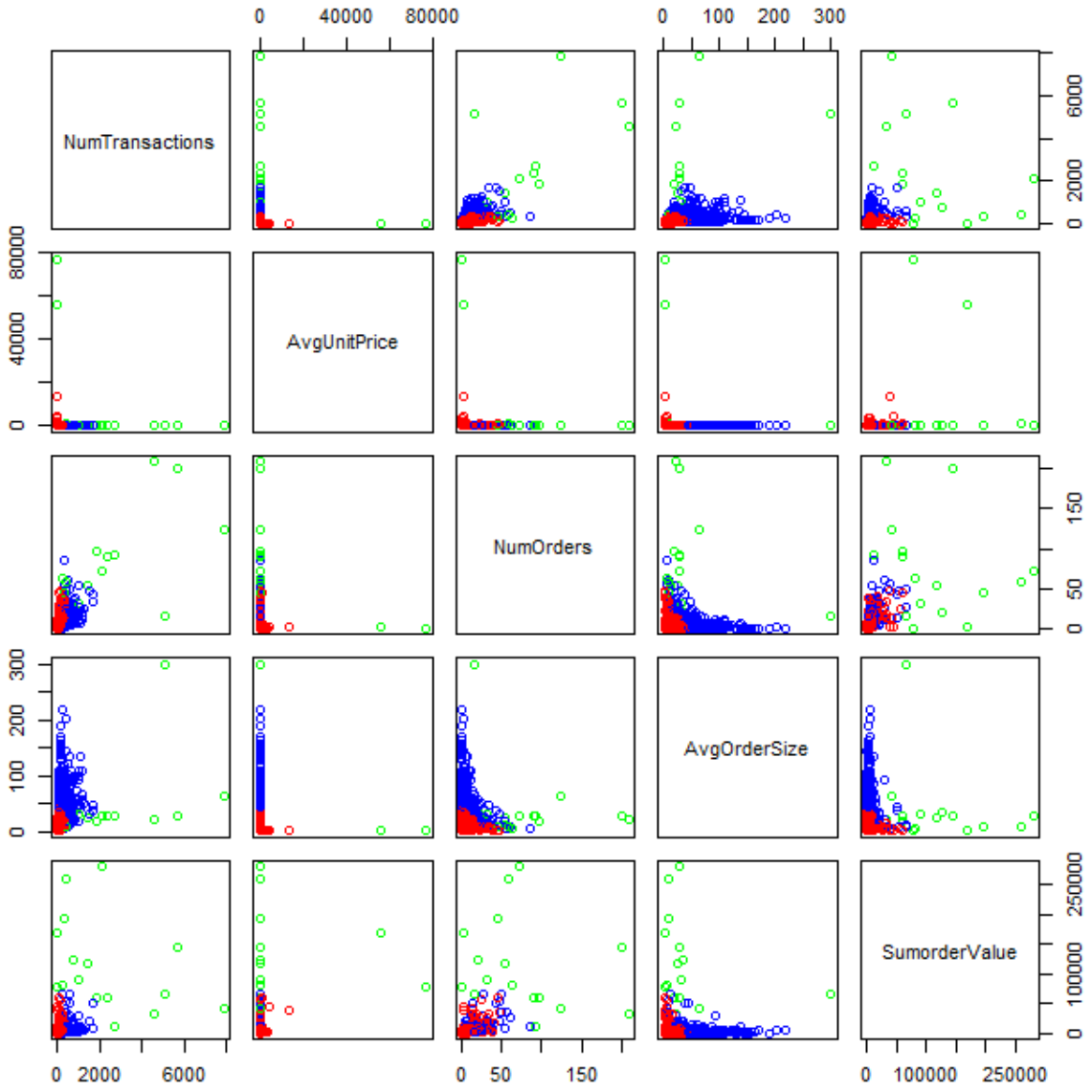
Table 5

Size and Ratio of each Cluster

	Size	Ratio	Color
Cluster 1	3552	81.88%	Red
Cluster 2	770	17.75%	Blue
cluster 3	16	0.37%	Green

Figure 7 shows each feature against one another with the observation colored according to their cluster. This is done to demonstrate whether meaningful groups have been formed.

Figure 7



Notice that there are clear groups present in some of the plots. This supports the groupings created by the algorithm. Table 6 shows the summary statistics for each group. Further, figure 8 shows the boxplots of every group, demonstrating the summary statistics in comparison to each other.

Table 6

Summary Statistics of Customer Purchasing Habits Distinguished by Cluster

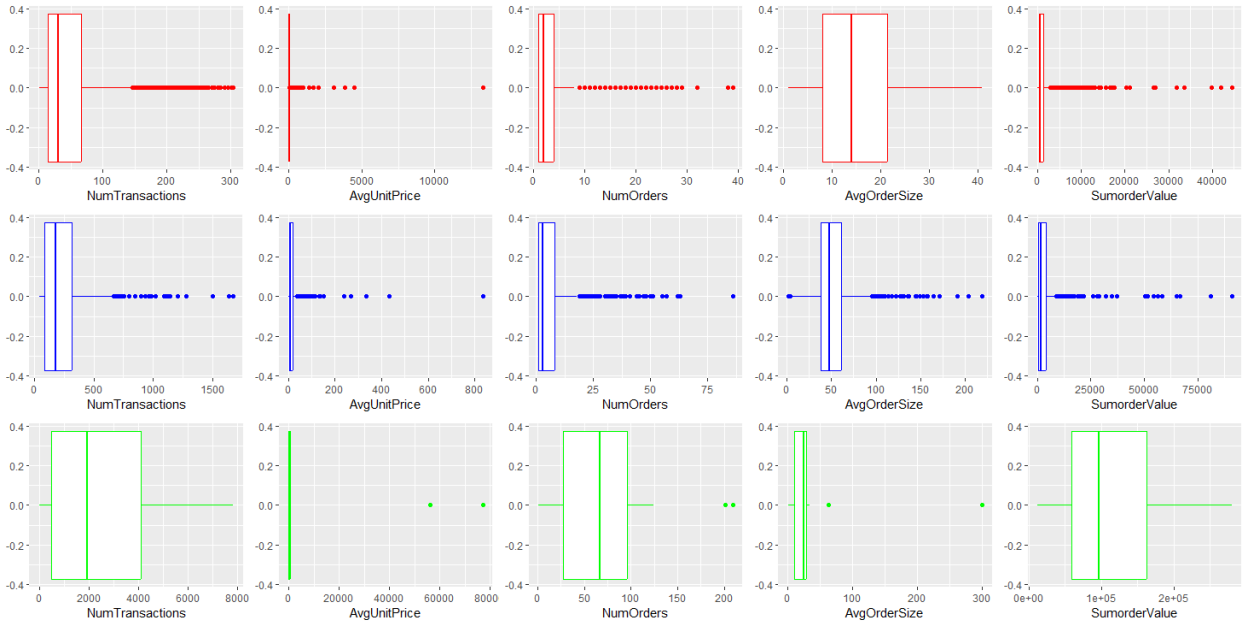
Cluster 1	NumTransactions	AvgUnitPrice	NumOrders	AvgOrderSize	SumorderValue
min	1.00	2.15	1.00	1.00	3.75
1st quarter	14.00	14.97	1.00	8.14	271.32
median	30.00	18.73	2.00	14.00	579.77
mean	49.07	42.12	3.37	15.41	1,185.89
3rd quarter	67.00	26.82	4.00	21.50	1,313.33
max	304.00	13,305.50	39.00	41.00	44,534.30

Cluster 2	NumTransactions	AvgUnitPrice	NumOrders	AvgOrderSize	SumorderValue
min	42.00	2.10	1.00	1.56	120.03
1st quarter	88.00	4.97	1.00	38.33	650.43
median	179.00	8.47	3.00	48.00	1,635.66
mean	239.99	15.94	7.02	52.20	3,905.56
3rd quarter	319.00	17.78	8.00	61.09	3,984.22
max	1,677.00	835.86	86.00	219.00	91,062.38

Cluster 3	NumTransactions	AvgUnitPrice	NumOrders	AvgOrderSize	SumorderValue
min	1.00	4.50	1.00	1.00	12,156.65
1st quarter	501.75	15.90	27.25	10.18	59,311.35
median	1,947.00	58.10	66.50	25.80	97,281.62
mean	2,506.21	9,644.83	77.86	42.35	116,986.85
3rd quarter	4,121.25	476.71	96.00	28.88	162,310.64
max	7,847.00	77,183.60	209.00	300.65	280,206.02

Figure 8

Boxpots of each cluster (group), where red is cluster 1, blue is cluster 2 and green is cluster 3



Now that the data is split into groups, these summary statistics are much more visually interpretable. Notice the feature SumorderValue. Cluster1 has the lowest average value of 1,185.89, cluster2 has the middle average value of 3,905.56 and cluster3 has the highest average value of 116,986.85. Further, the metrics between the clusters contain interesting differences. Take for instance the feature AvgUnitPrice. Cluster1 has an average, 42.12, that is higher than cluster2's average, 15.94. Yet, cluster2 NumOrders and AvgOrderSize average values are larger than cluster1's, leading to the higher SumorderValue average.

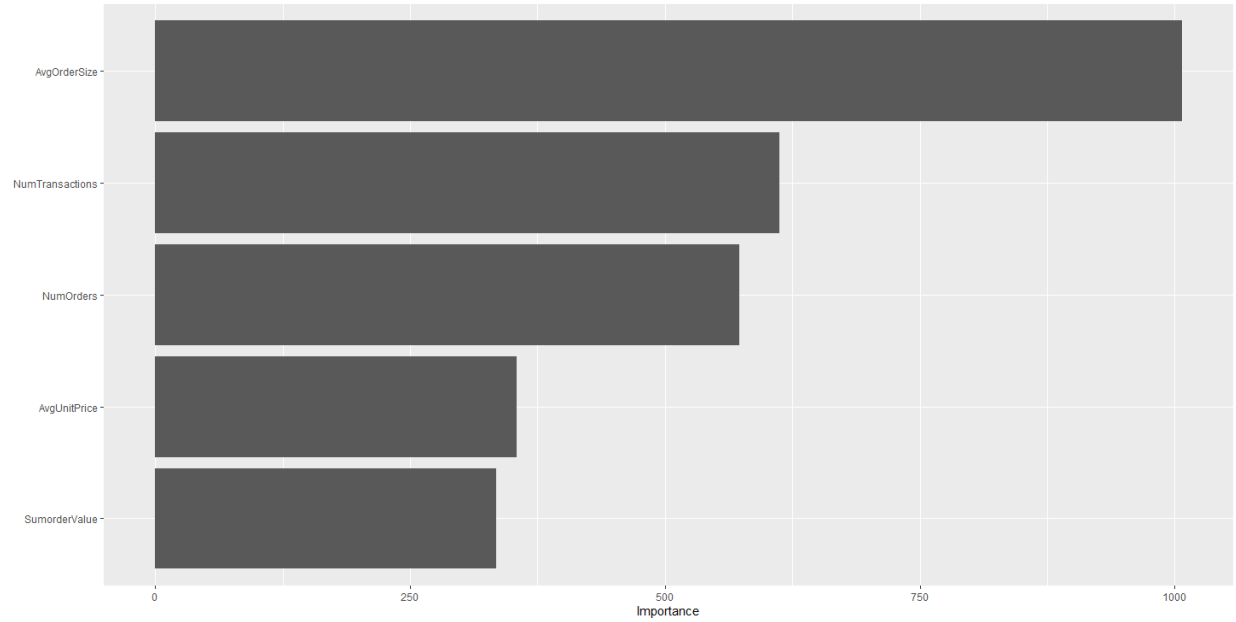
The purpose of this exercise is to demonstrate the results of, and viability of the k-means clustering algorithm applied to the customer data. As clear, interpretable results have been formed, I will move to the next stage of the project: CART analysis.

CART Classification

Using the groupings formed by k-means clustering, guided by the SumorderValuefeature, I determine that group 1 will be considered "Low Value Customers", group 2 will be considered "Moderate Value Customers" and group three will be considered "High Value Customers". The groups, Low, Moderate, or High Value, will be treated as classes by the CART algorithm. This means that the algorithm will be analyzing which customers belong to which classes, i.e groups, and why. The algorithm will produce a decision tree to predict which customers belong to which classes. Figure 1 shows this decision tree. The CART analysis reveals which features are of most importance in determining classification. Figure 9 shows the relative impact of each (note that 'importance' is an arbitrary metric used by the VIP library to compare features)

Figure 9

Relative Importance of Features



Notice that the three most important features, AvgOrderSize, NumTransactions, and NumOrders, are the same features appearing in the decision tree. This suggests that they are the first metrics that should be examined.

At this point, I have produced a model that, as validated by the test data, reasonably determines which class the customers belong to. Now, we can apply this to our actual scenario and create guidelines for the business to follow.

Analysis of Groups

The CART analysis suggests that AvgOrderSize, NumTransactions, and NumOrders are the most impactful features. So, I will explore the differences between them. Table 7 presents the average of each of these features for convenient comparison

Table 7

Averages of Most Impactful Features Distinguished by Cluster

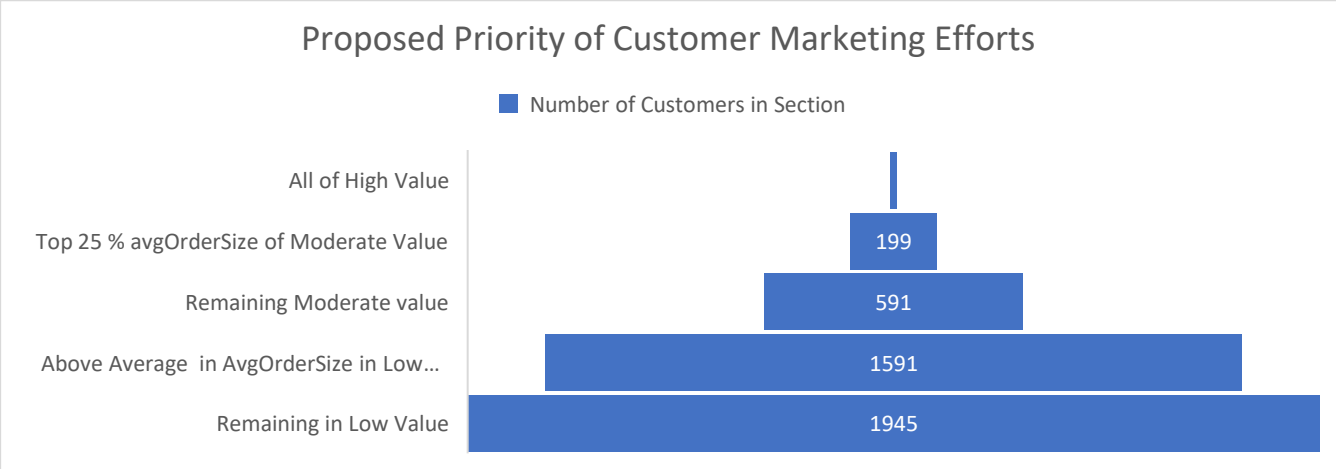
	Mean of AvgOrderSize	Mean of NumTransactions	Mean of NumOrders
Cluster 1	15.41	49.07	3.37
Cluster 2	52.20	239.99	7.02
Cluster 3	42.35	2,506.21	77.86

Notice the generally linear increase between clusters that is broken by the AvgOrderSize of cluster 3. This suggests that across customer classes, there is a general limit to how large orders will be. In comparison, the overal number of transactions, and the overall number of orders both continue to increase as customer class increases, suggesting that customer class is strongly impacted by the frequency the customer makes separate order of a certain size. This lays the groundwork for a deeper, predictive analysis of customer behavior for each group. This potential will be further explored in the discussion.

Finally, I will discuss the broad purpose of customer classification and how the business may use this knowledge. There are two perspectives, or angles, that this information may be used for. The first perspective is that knowing which customers bring the most value allows the business to properly cater marketing and sales efforts to maintain and grow relationships with those premium customers. Knowing what factors separate a high vs moderate value customer may help to detect which moderate value customers have the potential to become high value. Considering the most impactful feature, AvgOrderSize, in figure 10 I propose the following hierarchy for allocating targeted sales efforts.

Figure 10

Proposed Priority of Groups and Group Subsets



This proposal allows the business to first address the most important customers, followed by the top 25 % of moderate value customers, followed by the remaining moderate value customers, the top 50 % of low value customers and any leftover resources may be directed to the remaining low value customers.

Before moving on to an overall discussion of the project, I will summarize the results and the steps the business may take. The surface level analysis revealed which products are most in demand and a broad model for when products are most demanded. This may be used to inform when the business places inventory orders and which products populate the orders. Next, it was demonstrated that the majority of the business's orders come from UK based customers. This tells us that the business should consider expanding to the international market. Three countries, Germany, France, and Ireland, were highlighted as being the best starting places for this market expansion. Lastly, the current customers were classified into low, moderate, and high value. This resulted in a model to classify new customers. Further, the customer classification was used to create a priority model that should inform the business's marketing and sales efforts.

Discussion

The goal of this project was to perform exploratory analysis and machine learning modeling on the UCI Machine Learning Repository Online Retail dataset to produce a classification model and meaningful insights the business may use to inform their operation decisions. This was successfully executed and insights have been created concerning market expansion, inventory management, and sales/marketing efforts.

Further Analysis Opportunities

However, this was not a comprehensive study of the data. Rather, this project is only a piece of a larger understanding of this data. Further, in executing this project, many different angles of analysis and points of deeper exploration have been revealed. Here, I will give an overview of where future analysis should be done.

The most significant avenue that has not been explored is analyzing which products are often purchased together. This type of analysis—market basket analysis—may be performed to create assertions about which products are likely to lead to the purchase of other products. This would then allow the business to target the advertising of specific products to customers most likely to purchase them.

Related to this is an analysis of the likelihood of specific products being purchased at different times of the year. By determining when demand is highest for a given product, we can recommend targeted advertising of that product when it is most likely to be purchased.

Finally, turning to the customer purchasing habits, using the features I created from the original data, a deeper, predictive analysis may be done on each metric, starting with the highest impact metrics. This would allow the business to make specific predictions about customer behavior. For instance, we may be able to predict how many orders and the size of the orders a given customer is likely to make. This would be part of a larger study into the business's expected revenue.

Conclusion

While this was by no means a comprehensive study, it nevertheless resulted in many impactful insights the business may use to efficiently direct their efforts. Inventory manegement, largescale market direction, and nuanced, targeted marketing efforts all can be guided by the results of this project. Further, the extensive understanding and preparation of the dataset provides the foundation for future projects that will continue to inform the business's most efficient processes. This analysis represents invaluable information that should be applied by the business and will ongoingly increase the effectiveness of their operations.

`

References

Online Retail. (2015). UCI Machine Learning Repository. [Data set] https://doi.org/10.24432/C5BW33.