Data Analysis for 7Lytix

Max Allinger-Csollich

July 2021

1 Introduction

The data was provided via `https://www.kaggle.com/mashlyn/online-retail-ii-uci`. It contains sales of an online retail store from December 2009 up until the end of November 2011. These are sorted as over 1.000.000 individual product purchases, providing the following information:

- Invoice - The Number of the current invoice, multiple products can be sold with one invoice
- Stock Code - A tracker for the individual product
- Description - Name of the product, this entry can last over several list elements, if the function *split(sep=',')* is used to convert *.csv* file into a python list
- Quantity - Number of instances the product was bought
- Invoice Date - Date and time of the invoice, constant for all products of the same invoice
- Price - price of this product in British pounds £
- Customer ID - Assigned to each unique customer, allows tracing them over several months
- Country - home country of the customer, the majority is from the United Kingdom, as shown in figure 1

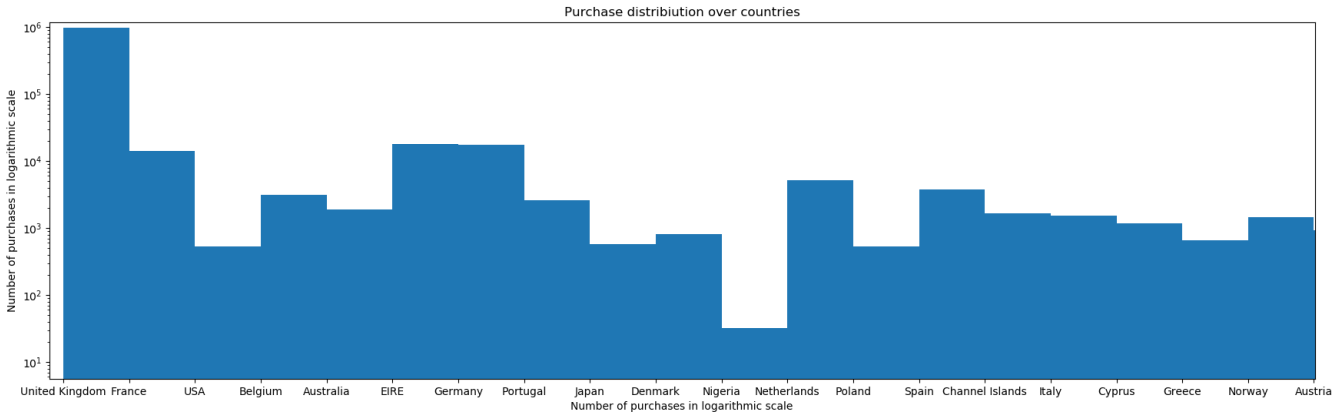

Figure 1: Amount of purchases per country. Note the logarithmic scale. The United Kingdom is home to more than one hundred times the next largest recipient, Ireland. Overall 42 different countries are being tracked with 19 being shown in this graphic. No other country comes close to France, Ireland or Germany.

1.1 Limitations

The data set has several limitations which had to be accounted for, these were both of statistical nature and of systematic errors in the data set, like missing or incorrect entries.

1.1.1 UK dominance

Customers from the United Kingdom are responsible for nearly 92% off of all purchases which takes a lot of statistical weight of all other countries. Nevertheless, country dependence could be analyzed, but was not done for this project.

1.1.2 Missing and Fraudulent Data Points

The vast majority of all entries was in perfect condition, sometimes containing only minor inconvenience, like additional characters, such as " (quotation marks), . (points) or additional spaces. These were stripped via the *cut_waste* function. Another problem appeared in the form of empty entries, or missing values. While negative costs were first suspected to represent refunds, these only showed up three separate times and were as such suspected to be a rare occurrence. Thus, the *customer* class only appends a new purchase, if it provides a positive price.
Approximately 8500 purchases were linked to the same customer, a rather high number, considering the 730 days of recording, this would represent more than 11 product bought each day. After checking the name of the customer, it had length 0, thus representing all customers who did not get assigned a customer number. Thus all those products were scrapped when it came to looking for returning customers for section 3.

2 Expected Income

Many external factors can influence retail revenue. The data set falls at the end of the global financial crisis which started in 2007 and right into the euro crisis from 2009 to 2011. Since the latter did not effect the UK that much it is not considered in this analysis. The chosen approach is of purely statistical nature to provide a good overview of the expected revenue over the coarse of the year. Should the total revenue have increased over the last decade, one can use it as a baseline for relative proportions based on relative total revenue comparing the current year(s) with the given data.

2.1 Method

A Machine Learning approach was considered but the more secure and stable statistical approach was chosen. For this the median daily income was calculated for each calendar week. The median was chosen to eliminate irregular highs of lows and smooth the development over the year. The resulting income is shown in figure 2 as the dashed line. Afterwards a rolling average is calculated over intervals of four weeks to provide cornerstones for the fitted cubic spline. The resulting function gives a model to predict the expected income on any given day. It is shown as the solid line in figure 2.

2.2 Interpretation and Application

The resulting function is given as a function $I\,[£] = I(day\,[\#week])$. It can predict the expected income for a single day, considering that a week provides eight sub steps. As the week starts on a Monday, the expected income from Wednesday to next Tuesday would be

$$\sum_{n=2}^{8} I\left(w + \frac{n}{8}\right)$$

considering the first Wednesday is in calendar week w.
The graph shows a clear behavior over the year with a peak in income in the weeks before Christmas. This is to be expected for an online retail store. The second, smaller peak appears to be around the middle of May with sales falling of until the end of July. Note that Easter in the two relevant years fell on the 4^{th} of April 2010 and on the 24^{th} of April 2011. These increases might also influence the rise of sales in the spring.
As shown in section 3 and figure 4, the increase in Christmas sales marks a large portion of irregular buyers that do not regularly use this retailer's services.

3 Returning Customers

The second task required a model to estimate the Chance that a given customer would return the following quarter after making a purchase. This is done via two separate analyses, one for the time of purchase and one for the purchased product and purchased product type.

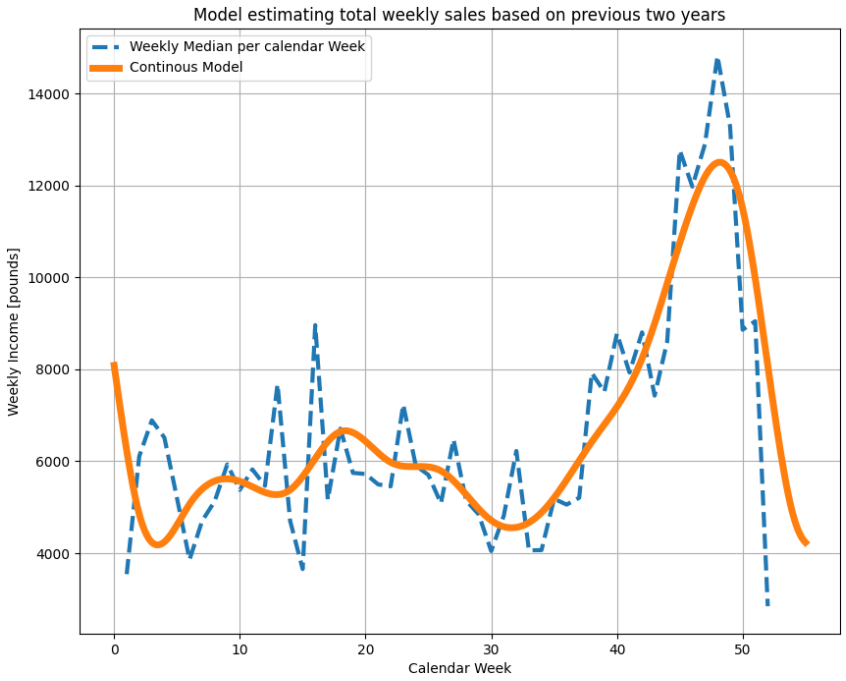

Figure 2: Spline fit and weekly median for income per week. The increase in sales before Christmas is the most obvious feature with the late spring/early Summer also showing a clear increase in income. The lowest performing times are the ends of January and August respectively.

3.1 Method

Every customer's invoices were analyzed and compared to other invoices of the same customer. An invoice was stored as a returning invoice, if the customer made any other purchase in the following quarter. If no such invoice is found, the first is stored as a non returning-invoice.
These invoices are then dissected either based on all contained products or the calendar week they were purchased in. For those two parameters, ratios were compared for each individual element, comparing those that returned the next quarter to the total number of this element (product or calendar week).

3.2 Returning Customers based on product

Several products have 100% return rate in the next quarter, the most purchased being the "*CANDY PINK FRENCH BOUDOIR CLOCK*", purchased 118 individual times, each time, the customer returned the next time. The "*DOT-COM POSTAGE*" was purchased 1333 times with a return of the customer the next quarter and boasts with 92.7% also a very high return frequency. The most purchased product is the "*WHITE HANGING HEART T-LIGHT HOLDER*", with 4183 returning purchases and a 78.16% return chance.
On the other side of the spectrum are even less frequently purchased overall, with the "*SWEETHEART BIRD HOUSE*" being purchased 95 times and returning only once in the next quarter, resulting in a 1.58% return chance. The product purchased the most without the customer returning is the "*JUMBO BAG PAISLEY PARK*", purchased a total of 342 times, with no customer returning the next quarter.
Low return rates can have multiple reasons, a small sample size being one of them. Other reasons might be either an above average suitability for the Christmas season, which features a very low return chance overall (see section 3.3) or maybe a poor quality of the product or products purchased simultaneously. Since the products themselves are known only by name, no conclusions can be drawn here. It may only be noted, that "overly Christmas names" do not appear to be present at an above average rate, when scanning the list of products with low return rates.

Since several products are of a very similar type, these types were analyzed via the same principle, with product names having to contain individual words. For example, 40 separate products contain the word "*LUNCH*", the corresponding ratio would be 79.1% return rate with 22362 customers returning after purchasing an item with "*LUNCH*" in its name. As another example "*CHRISTMAS*" in an article's name had 20457 customers returning,

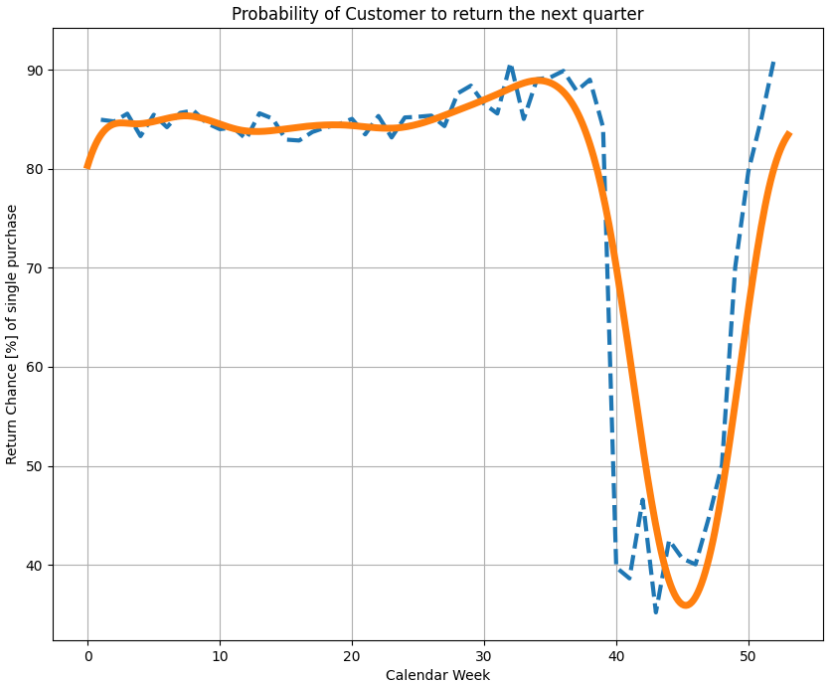

Figure 3: Relative chance that a customer returns the next quarter. Depicted is both the weekly ratio and the rolling average over four weeks fitted into a continuous spline. Notable is the sharp drop of in the early December and November, as a result of the high number of single time customers before Christmas.

resulting in a 60.5% return chance. The Appendix contains two tables, with table 1 containing the return chance for the 50 products with the highest absolute number of returning customers, table 2 containing the 50 highest returning words.

3.3 Returning Customers over the year

For this it is to be noted, that a customer, who made two purchases in January, one in February, two in March and one in April will show up 5 times in the plot, once for each week in the first quarter, as it cannot be said, which of these purchases "caused" the return. Should the customer return multiple times in the next quarter, every purchase in the previous quarter only counts once for return chance, independently of the number of returning customers. Figures 3 and 4 show the development and numerical spline generated similarly to section 2.1. Figure 3 shows the relative chance of return with a clear drop of during November and December, the same time figure 2 shows the highest amount of sales. This was to be expected, as this time shows a high amount of one time customers that only shop for Christmas presents and ordeals. Similarly, the time around the 35^{th} calendar week shows the highest amount of returning customers, both because the time features a generally low amount of income, thus presumably less customers, and because nearly every regular buyer will purchase items during the Christmas months.

Likewise, figure 4 shows the absolute number of customers each week that returned the next quarter. The shape more generally resembles the income plot in figure 2 for the spring and summer months. The high return rate in these months leads to the conclusion that these customers are regular customers that purchase a very consistent amount among each other, thus giving both curves a proportional shape.

The fact that the drop of for the Christmas purchases is still notably, but less prominent, thus supporting the assumption, that the increase in revenue comes mainly in the form of one time Christmas customers.

4 Basket Recommendations

This numerically simple but computationally exhausting task asked for recommendations for further products based on products already in the shopping cart. To do this, a grid was generated with one row for each individual product. Each entry in the row contains the ratio of invoices that contained both the requested product as well as any other given product.

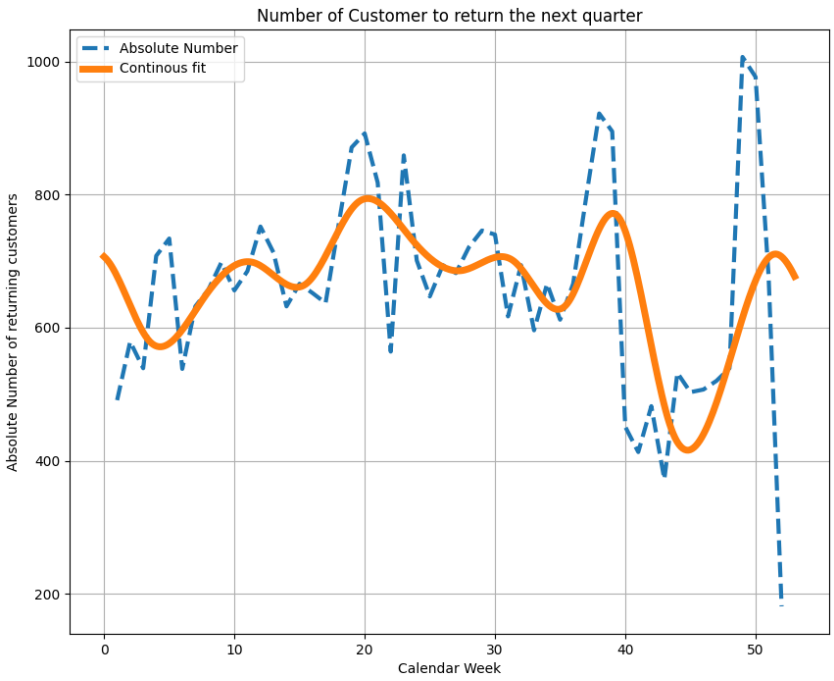

Figure 4: Absolute number of customers, returning the next quarter. Depicted are both the weekly ratio and the rolling average over four weeks fitted into a continuous spline. The mostly constant amount of customers leads to the suggestion of a constant base of regular customers who return each quarter, approximately 700 in size. Taking weeks 5-25 of figure 3 this leads to the assumption, that most of these 700 return at least once per quarter, most likely several times

4.1 Method

All invoices were sorted into a single list. For each product, this long list was then scanned once for each product A with all invoices containing the individual product A noted down as L_A. Afterwards, each invoice in L_A was scanned again once for every other product B, noting the individual ratios between the number of invoices N_{AB}, containing both products A and B, and the total number of invoices N_A, containing at least product A. This ratio N_{AB}/N_A was noted for every combination of products possible, creating a $(n \times n)$ matrix M with n being the total number of products. The diagonal elements represent the correlation of a product with "itself", which would always be 1. Thus this point was used to mark down the total number of times, product A was purchased. The number is written as negative, meaning the minimum of any row always gives the total number of invoices containing the product. This is useful as a reference, as a product which was only purchased once or twice does not provide significant statistical weight.
This matrix is then saved into a *pickle* file which can be accessed via the special *search_basket* function, provided in the *Retail* module. The result of the function is an array with the top five corresponding products.
Please note, that the full *basket.p* file cannot be uploaded onto GitHub, as it is over 200 MB in size. Should you be unable to generate it yourself, please contact the author to receive a OneDrive Link to the file.

4.2 Analysis

After some testing, the function worked perfectly, there is not much special analysis to do, as each individual product gives a different result. As is to be expected, more frequently sold products have lower maximum correspondence in the basket, as they are spread across more than 1000 invoices.
A notable exception of this would be the " *SPACEBOY LUNCH BOX*", being bought a total of 1382 times, 60% of which also featured the "*DOLLY GIRL LUNCH BOX*". The latter was purchased 1353 times itself. The second highest correlation for both of these products is the "*LUNCH BOX I LOVE LONDON*" with just under 30% correlation for both of them.

Product	# returning customers	return chance [%]
PACK OF 72 SKULL CAKE CASES	1114	77.96
HAND OVER THE CHOCOLATE SIGN	1118	80.03
HANGING HEART JAR T-LIGHT HOLDER	1123	80.33
JUMBO BAG WOODLAND ANIMALS	1126	81.36
72 SWEETHEART FAIRY CAKE CASES	1144	79.22
REX CASH+CARRY JUMBO SHOPPER	1160	64.48
RECIPE BOX PANTRY YELLOW DESIGN	1169	71.59
RED TOADSTOOL LED NIGHT LIGHT	1169	78.04
PLEASE ONE PERSON METAL SIGN	1179	75.53
PACK OF 72 RETRO SPOT CAKE CASES	1199	84.5
PACK OF 72 RETROSPOT CAKE CASES	1200	73.17
HANGING HEART ZINC T-LIGHT HOLDER	1208	81.79
GIN + TONIC DIET METAL SIGN	1222	78.69
SWEETHEART CERAMIC TRINKET BOX	1223	80.2
BAKING SET 9 PIECE RETROSPOT	1226	67.07
SMALL POPCORN HOLDER	1231	82.9
VINTAGE SNAP CARDS	1245	70.74
JUMBO BAG PINK VINTAGE PAISLEY	1247	78.72
WOODLAND CHARLOTTE BAG	1259	81.44
BATHROOM METAL SIGN	1269	80.78
6 RIBBONS RUSTIC CHARM	1271	75.39
RED HANGING HEART T-LIGHT HOLDER	1274	75.83
RECYCLING BAG RETROSPOT	1282	82.55
POSTAGE	1301	72.28
DOTCOM POSTAGE	1333	92.7
LOVE BUILDING BLOCK WORD	1343	81.05
HEART OF WICKER LARGE	1344	71.76
LUNCH BAG RED RETROSPOT	1380	76.24
VICTORIAN GLASS HANGING T-LIGHT	1419	80.03
NATURAL SLATE HEART CHALKBOARD	1425	73.61
JUMBO BAG BAROQUE BLACK WHITE	1449	79.57
JUMBO BAG STRAWBERRY	1478	81.61
PACK OF 60 PINK PAISLEY CAKE CASES	1490	77.48
LUNCH BAG SPACEBOY DESIGN	1514	75.97
LUNCH BAG WOODLAND	1538	80.27
HEART OF WICKER SMALL	1561	75.96
LUNCH BAG CARS BLUE	1577	78.42
WOODEN PICTURE FRAME WHITE FINISH	1588	81.19
WOODEN FRAME ANTIQUE WHITE	1596	80.04
HOME BUILDING BLOCK WORD	1647	81.05
60 TEATIME FAIRY CAKE CASES	1647	78.21
JUMBO SHOPPER VINTAGE RED PAISLEY	1664	78.53
KEY FOB	1788	74.31
LUNCH BAG BLACK SKULL	1830	79.39
JUMBO STORAGE BAG SUKI	1844	80.63
ASSORTED COLOUR BIRD ORNAMENT	1887	72.8
STRAWBERRY CERAMIC TRINKET BOX	2042	82.57
PARTY BUNTING	2083	82.5
JUMBO BAG RED RETROSPOT	2457	77.53
REGENCY CAKESTAND 3 TIER	3156	78.22
WHITE HANGING HEART T-LIGHT HOLDER	4183	78.16

Table 1: Return chance per product, the second column gives the absolute number of customers returning the next quarter after having purchased the specific product.

Word	# returning customers	return chance [%]
CARD	20191	74.84
SMALL	20341	76.66
CHRISTMAS	20457	60.49
COL	20458	75.4
DEC	21173	75.16
PACK	21750	74.79
PAPER	21874	78.42
STR	22170	79.69
LUNCH	22362	79.08
JUMBO	24111	77.92
T-LIGHT	24127	76.23
HANGING	24798	78.05
HOLDER	25389	76.63
HOLD	25408	76.63
CAR	26524	74.64
LAS	26923	75.72
TEA	26929	77.71
CAN	28564	75.46
ANT	28723	74.29
OLD	29364	76.45
BLUE	30154	78.09
BLU	30173	78.1
METAL	31880	77.93
VINTAGE	32719	72.58
VINT	32746	72.58
WOOD	33188	74.9
LIGHT	33370	75.91
GIN	34137	77.23
ASS	35462	76.52
WHITE	35789	78.64
DESIGN	35829	74.47
DES	36930	74.68
TIN	37000	75.9
ALL	37296	76.14
CAKE	37957	77.07
TAG	39567	73.46
RETROSPOT	39711	75.91
BOX	40157	75.46
PINK	46342	79.15
PIN	49928	78.47
RETRO	50076	77.41
SIGN	61348	76.33
HEART	65139	75.75
BAG	67738	78.88
RED	67927	78.32
EAR	69880	75.79
AND	71773	75.88
SPOT	74760	79.52
SET	75016	74.65
ART	79277	76.02
POT	83812	79.16

Table 2: Return chance if the product name contains a specific word. Please not, that this list also contains filler words like "*AND*".