Report
19-02-20

E-commerce, a look behind the hidden curtain

By Mohamed Khanafer

Table of Content

Scenario description

With 40% of worldwide Internet users (which amounts to a billion users) having purchased goods online, the role of e-commerce in the world economy has become very extensive. Not only has online shopping allowed stores to increase sales, but it has also allowed them to break the boundaries they were once limited to. With these far-reaching impacts and those continuously growing numbers, it has become a must for anyone involved in business to get a core understanding of the data underlaying e-commerce businesses.

However, getting hands on the data of an e-commerce only store is not an easy process. Most of this information is considered private and sharing it could constitute a disadvantage for the business. But, luckily, the UCI Machine Learning Repository has made publicly available a dataset containing actual transactions from 2010 and 2011 for an Online Retail business based in the United Kingdom. The main products sold by the company are unique all-occasion giftware. Also, many customers of this e-commerce are wholesalers.

This opportunity gives us an eye into the hidden e-commerce numbers and allows us to ask ourselves some questions that we will answer through an elaborate analysis.

The Analysis' goals

With the current day advances in the field of Data Science, many could be tempted to directly jump into applying predictive analytics to get insights from the available data for a possible future horizon. However, the basic data understanding and exploration is the crucial step that should not be overlooked. And on the contrary, allocating the most resources here would better guide decision makers and stakeholders. This in-depth understanding of the data will allow the ecommerce to experience impactful growth and also to implement expansion strategies.

Our analysis here highlights how this research for information and hypothesis from the data analysis phase could already answer some crucial unknowns the business requires to be informed of.

The goal of the analysis is thus to get a general overview on some operation and performance metrics of a fully online retailer based in one country. This will help the reader assess the potential impact of the e-commerce sphere on businesses around the world.

We will list some hypothesis that will set the path for our exploration and we will try to confirm or challenge them in the analysis that follows.

Hypotheses as the analysis guidance

The different hypothesis addressed (directly or indirectly) are the following:

- The impact on revenues from online shopping is very significant. Indeed, with reduced costs and a wider market access, revenues for an online shop could be very high. Is that the case here?

- Most e-commerce platforms have access to markets beyond the borders of their country of origin. Online shopping could have thus brought geographic diversification to the online retailer. To what extent can this be shown here?

- With no physical locations and a fully online service, the ecommerce could thus take advantage of timing. Do we see patterns or trends for the analyzed business?

- For most e-commerce businesses, growth is an important variable. This could be done through various ways (marketing, promotions, etcs.). How relevant would this be for the analyzed e-commerce?

A deep dive into the Analysis

Jupyter notebook content

As this report extends from the detailed analysis in the attached Jupyter Notebook, I list here the content of that file to be able to easily link the 2 documents. In this section of the report, I explain the reasoning behind the analysis as well as the relevant observations to keep in mind.

The attached Jupyter notebook is organized as follows:

1. PySpark environment setup
2. Data source and Spark data abstraction (DataFrame) setup
3. Data set metadata analysis
    A. An Introduction to the Dataset
    B. Display schema and size of the DataFrame
    C. Get one or multiple random samples from the data set
    D. Data entities, metrics and dimensions
    E. Column categorization
4. Time Data Expansion
5. Columns groups basic profiling to better understand our data set

A. Timing related columns basic profiling
B. Sales related columns basic profiling
C. Customer related columns
D. A Filtered dataset

6. Getting some Insights to make the e-commerce grow
A. Revenues and number of orders, how good are they?
B. When do people shop?
C. What does customers analysis show?

Before jumping into researching answers to our hypotheses and other insights we may find, a basic understanding of the data on hand is necessary.

Dataset Description

As mentioned above, the used dataset comes from the UCI Machine Learning Repository (it can be found here: https://archive.ics.uci.edu/ml/datasets/Online+Retail+II)

This dataset contains all the transactions for an Online Retail business registered in the United Kingdom. The transactions took place from the 1st of December 2009 to the 9th of December 2011. The main products sold by the company are unique all-occasion giftware. The dataset contains 541,909 rows and 8 columns.

Understanding the columns in the dataset and their descriptions is crucial for understanding the flow in the analysis. Here is a description of the variables:

InvoiceNo: a nominal Invoice number. It is a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.

StockCode: the nominal product (item) code. A 5-digit integral number uniquely assigned to each distinct product.

Description: the nominal product (item) name.

Quantity: the quantities of each product (item) per transaction (numeric).

InvoiceDate: the invoice date and time (numeric). The day and time when a transaction was generated.

UnitPrice: the unit price (numeric). Product price per unit (in sterling pound).

CustomerID: a nominal customer number. A 5-digit integral number uniquely assigned to each customer.

Country: the nominal country name. The name of the country where a customer resides.

These columns can be categorized as follows:

- Timing related columns: InvoiceDate
- Sales related columns: InvoiceNo, StockCode, Description, Quantity, UnitPrice
- Customer related columns: CustomerID, Country

And also the different entities, metrics, and dimensions could be established as follows:

- Entities: Sales transactions (main one measured - facts), InvoiceNo (dimension), CustomerID (dimension)
- Metrics: Quantity, InvoiceDate, UnitPrice
- Dimensions: StockCode, Description, Country

A note on Data Profiling

The goal of the data profiling is to get more familiar with the data and it structure. Here, I made sure that the data was in a relevant format for the questions to be answered.
To be able to answer questions related to timing, I expanded on the available data from the column InvoiceData to include the columns: day, month, year and day of the week. These will allow us to get more insight on buying behaviors of customers.
Some irregularities like some missing values in the customerID or description of the products are missing, and this is taken into consideration later on. Also, an important observation is that the descriptions of the products are not unified and thus it would be hard to explore product-related questions.

The insights that could lead to the e-commerce's growth

Revenues Breakdown

The fundamental metrics that I believe is appropriate to start the analysis is the revenue breakdown. This gives us a general overview of the performance before allowing us to dive into more specific questions. To do this, I combined and derived numbers from the metrics columns defined above. I used them in combination to other columns like Country to make more sense of the numbers.

For a business that sells unique all-occasion giftware, a yearly 9,747,748 Sterling Pounds of revenue is quite impressive!
To better understand the source of this revenue, I broke it down into source by countries and as expected, the United Kingdom generated around 84% of the revenues, followed by the Netherlands, Ireland, Germany, and France which account for 10% of the revenues.
And we also find an interesting trend at the bottom of this classification: the shop sells to many countries in the Middle East, so even if this market constitutes a minimal amount of

sales, it could be interesting to keep in mind. The same observation holds for Southern Asian countries.

Also, this ranking is confirmed by the total number of orders, not only the sales. Indeed, almost 96% of total orders came from European countries.

Shopping Time Analysis

Next, to complement the geographical insights, it is relevant to analyze time related metrics.

From the analysis, we see that almost 55% of the transactions of the platform happen during the early afternoon (12:00-16:00). On a monthly scale, revenues flow in mainly during the last 4 months of the year (almost 50% of sales) with a peak in November. This could be because of the festive seasons.
On a weekly scale, sales tend to happen the most on Tuesdays and Thursdays (42%).
Overall, we can thus see that the timing variables are skewed whether it is on an hourly, weekly, or monthly scale.

The Lifetime Value of Customers

Like mentioned above, the customer information contained some missing data, thus I filtered the main data to perform a more coherent analysis.
After some transformations and combinations of information, I get very interesting insights: the platform had 4,372 customers during the year that on average spent 375 Sterling Pounds.
Averaging the number off orders, we see that customers on average bought 5 times form the store. But this information could be misleading. A more interesting metric would be to look at the number of times each customer bought from the store and then look at the distribution of this number. We get that 50% of customers bought 3 times or less with 30% only buying once. This shows a low conversion of customers into recurrent ones.

Conclusions

The main three take-aways of this detailed analysis based on initial hypotheses of this e-commerce shop would be:

- Out of the 35 countries it sold products, the store generated 80% of its revenue from the UK, with the rest coming mainly from Europe but has interesting clusters of countries it could consider targeting (namely the Middle East and South East Asia).
- The time patterns of sales (hourly, weekly, and monthly) are very skewed and the company could use this information to direct more sales to have a more balanced flow of transactions (this could be done through marketing).
- The company has a problem retaining customers and increasing their lifetime values with 50% of customers only buying 3 times or less and 30% buying only once. This metric could be improved by the implementation of various strategies and would ultimately lead to the growth of the online store.