Final Report

Tomas Costantino

Torrens University Australia

IDS201 – Introduction to Data Science

Ahsan Ali

22/08/2021

Table of Contents

Introduction

E-commerce is a growing sales channel that has become very popular lately, replacing the traditional face-to-face sale to a digital platform on which users can buy products from anywhere in the world. (Chen et al., 2021, p. 1). Indeed, this report defines a common problem in the field that needs to be tackled, then attempts to obtain insights about the issue by analysing a dataset of an e-commerce platform and describing the steps utilised to extract knowledge from such data to, finally, discuss about the results obtained from the analysis.

Data Science Problem and Dataset

In numerical terms, e-commerce has experienced a phenomenal growth in recent years, with a worldwide annual sales of 1,336 trillion USD in 2014, 2,842 trillion USD in 2018 and a projected figure of 4,878 trillion for 2021. (Ponce et al., 2020, p. 1). However, as the market grows and changes frequently, it produces a main unsatisfactory issue for customers and retailers, which is products not arriving on time. Hence, the challenge here is to understand not only e-commerce itself but also the effectiveness in its logistics services, as those services allow e-commerce retailers to reach further territories and thus become trusted by users.

Additionally, the analysis of the dataset is conducted to try to understand the "Why?" of the outcome, which is whether a product is on time or not. The dataset used for this study contains 10999 observations of 11 variables, and after renaming the variables to more conventional names for programming, they are listed as follows:

- block: the warehouse is divided into six blocks, A, B, C, D, E, F.
- shipment: products can be delivered via Ship, Flight, Road.
- care_calls: number of calls made to enquiry about the shipment.
- rating: rating from customers, 1 being the lowest (worst) and 5 the highest (best).
- cost: cost of product in US dollars.
- prior_purchases: number of prior purchases.
- importance: products can be of low, medium, and high importance.
- gender: gender of buyer, mapped as 1(Male) and 0(Female).
- discount: discount offered in US dollars.
- weight: weight of products in grams.
- on_time: this is the target variable to predict when building the model, a 1 indicates that the product reached on time, otherwise a 0.

Statistical Approach

The aim here is to extract knowledge from the data, which is achieved through the implementation of descriptive statistics and helpful plots. The methods utilised in the study are as follows:

- Pie chart: it basically allows to observe a portion of the total which lies in a particular category. The size of a pie slice shows the percentage for each category.
- Categorical plot: it condenses data series into an easily interpreted bar graph, where each bar represents a category, and the size of the bar represents the count of entries.
- Histogram: it works like categorical plots; however, it is used for quantitative data and the size of a bar displays the frequency of datapoints.
- Density plot: shows the distribution of variables over the target variable.
- Box plot: displays the distribution of numerical data and skewness through quartiles and averages.

These techniques were supportive to find similarities between products that reach their destination on time and those that do not.

Steps of the Analysis

Step one: Define questions about what is desired to discover, in this case:

- What is the success rate of the logistics system being analysed?
- Why do some products arrive late, and some others do not?
- What variables or features influence the outcome the most? In this case, the outcome being whether a product is on time.
- Is it possible to create a model that predicts, with high accuracy, whether a product is on time or not based on its features?

Step 2: Data acquisition. A public dataset is analysed which was acquired from the Kaggle platform right after researching about the e-commerce and its logistics.

Step 3: Data cleaning and preparation. Check for missing values and handle them.

Step 4: Exploratory Data Analysis (EDA). This is the most important step because it allows to extract insights from the data by identifying data types, doing univariate and bivariate analyses, identifying outliers, plotting distributions and so on.

Step 5: Data modelling. Repetitively apply Machine Learning techniques to identify a model that fits the issue and helps predict whether a product will arrive on time.

Results

To begin with, **Fig 1** shows that 59.67% of products did make it on time, whereas a slightly over 40% did not. In fact, it can be concluded that there exists a lack of effectiveness in e-commerce logistics services.

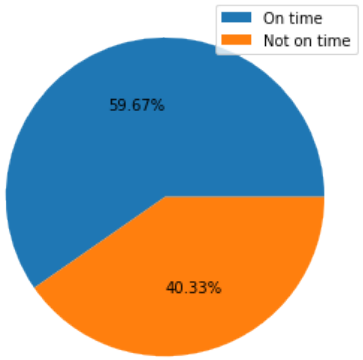

Fig 1

In addition, the categorical plots demonstrate that:

- Block "F" delivers most of the products, doubling the rest blocks. However, the efficiency of the blocks is not affected by the number of products. **Fig 2.**
- Most common shipment method is "Ship", and the type of shipment does not influence the outcome. **Fig 3.**
- The products managed by warehouse are mainly of "low" and "medium" importance although not altering the success of the logistics. **Fig 4.**

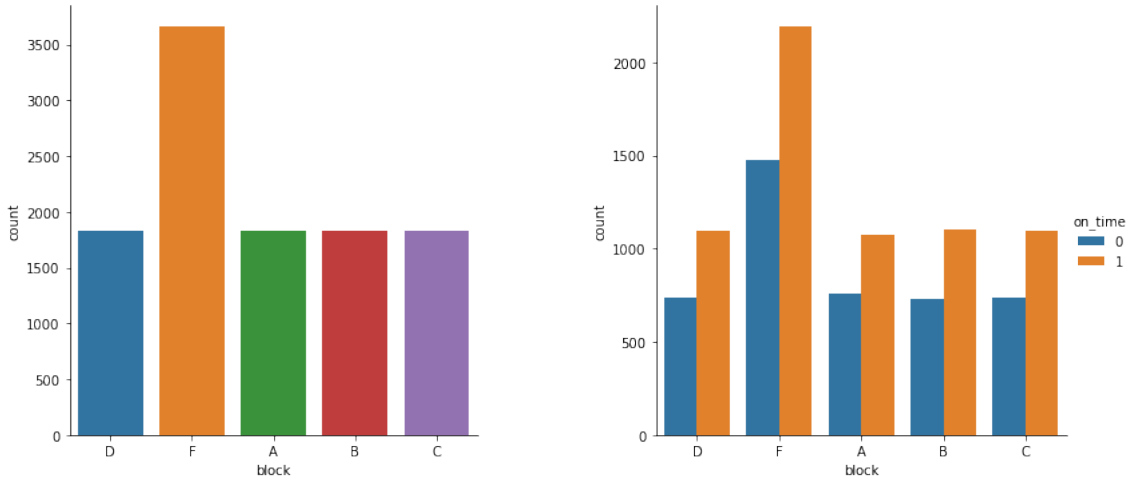

Fig 2

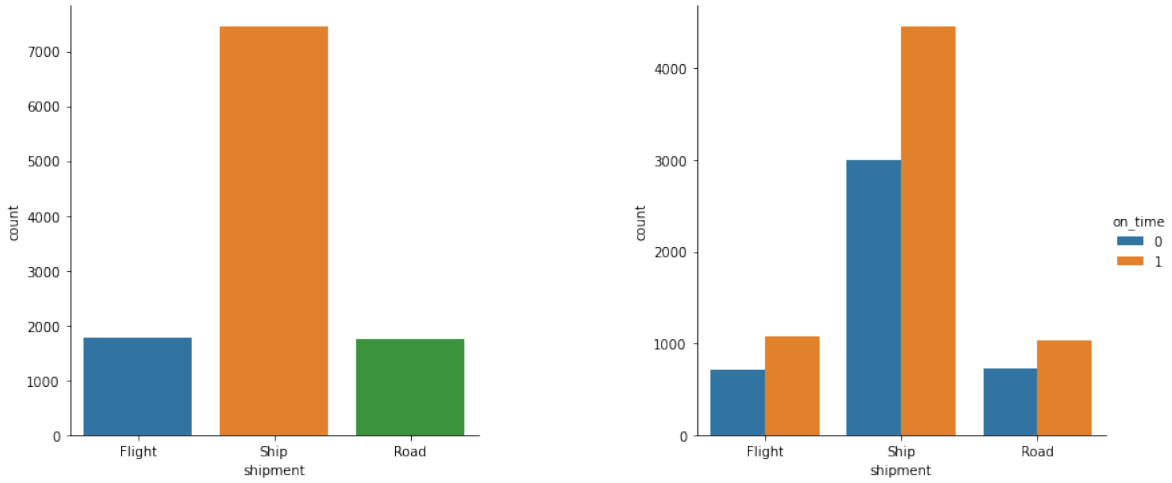

Fig 3

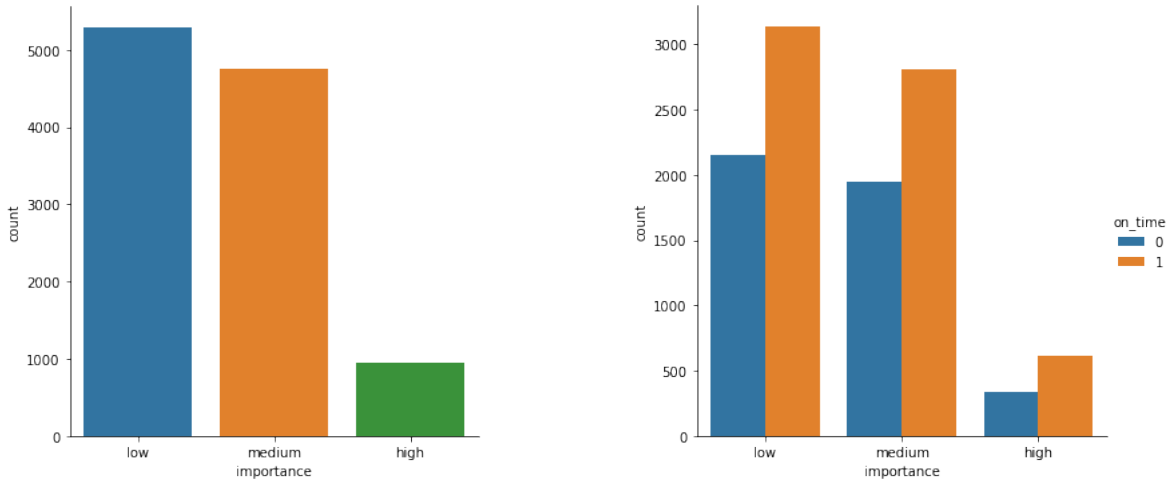

Fig 4

Then, moving on to the histograms, these show that:

- Most customers call between 3 and 5 times to enquire about their order.
- The prices are fairly distributed with a mean of 210 dollars per product.
- Nearly all users had bought between 2 and 5 products before the actual purchase recorded in the dataset.
- Discounts are usually not over 15 dollars.
- Most products weight between the ranges 1000-2000 and 4000-5500.

All these graphs are represented in **Fig 5**, referenced by variable names.

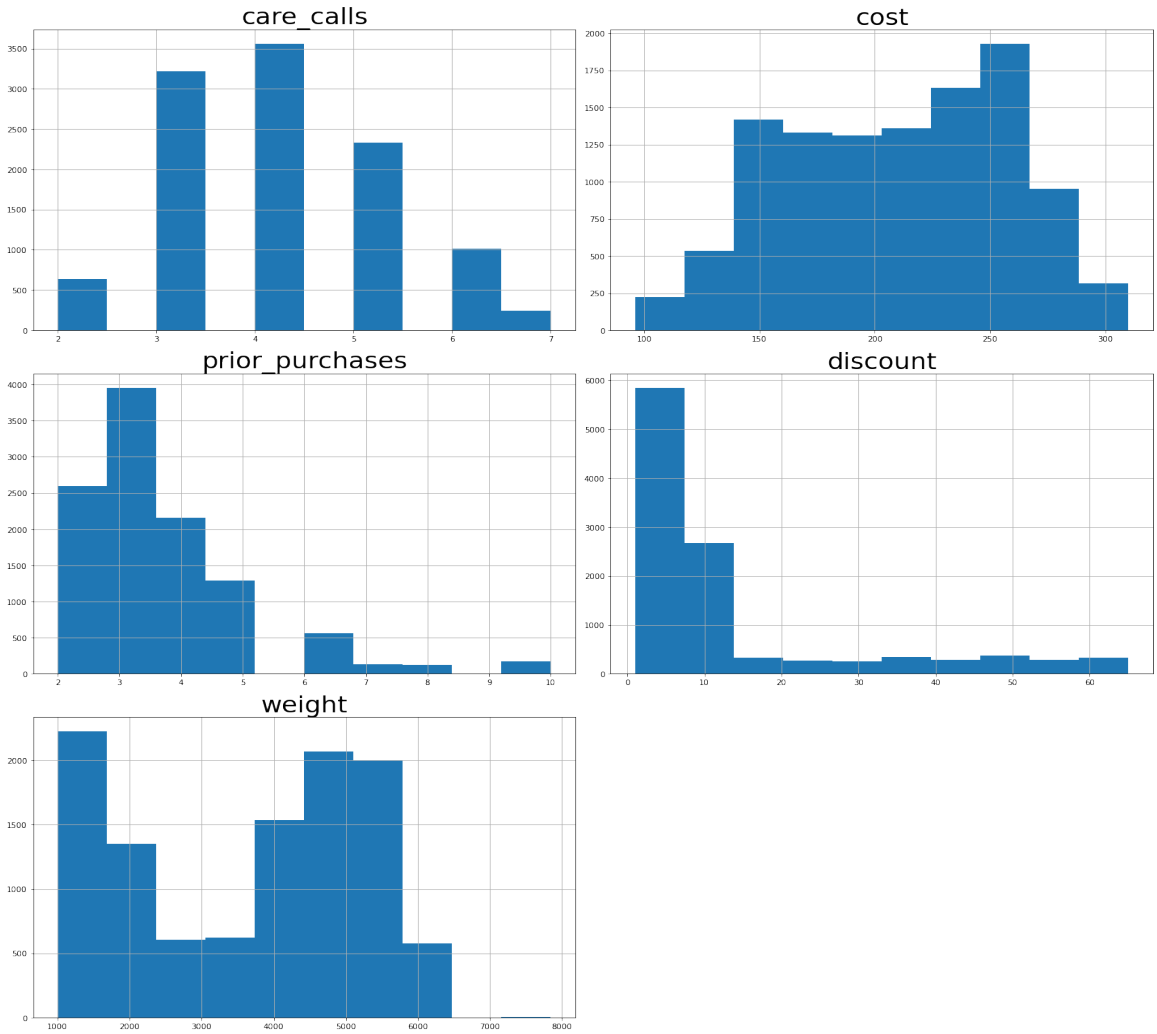

Fig 5

Furthermore, density graphs are plotted out to look at the distributions over the target.

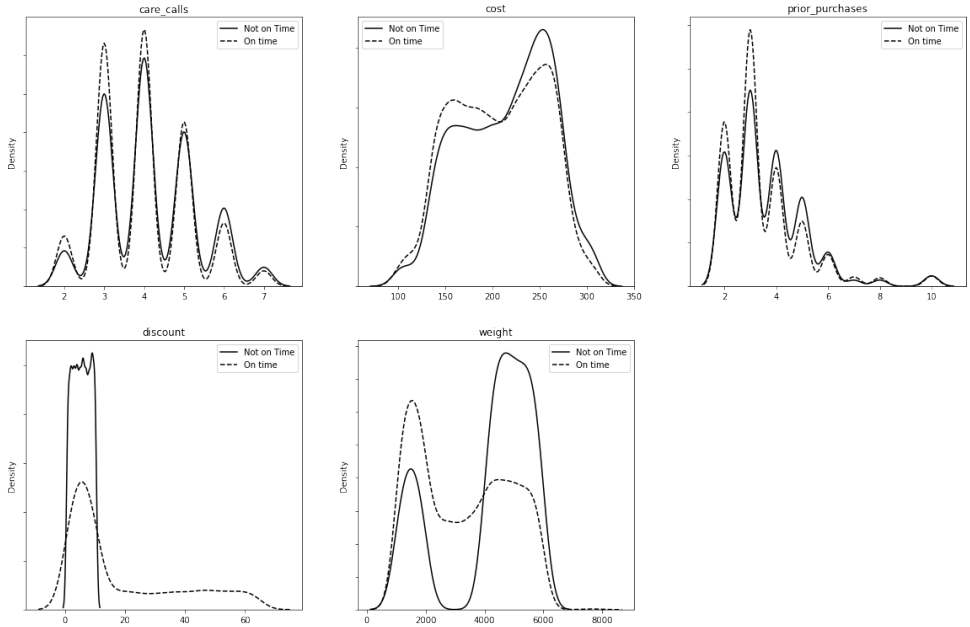

Fig 6

Finally, as the density plot displays that the means of “discount” and “weight” can be used as predictors for the outcome, a box plot is useful to prove it. Consequently, **Fig 7** discovers that:

- Heavy products ranging from 4000 to 6000 grams tend not to be on time whereas lighter ones are on time, regardless outliers.
- More products with high discounts reach their destination on the dot than those that have small price reductions.

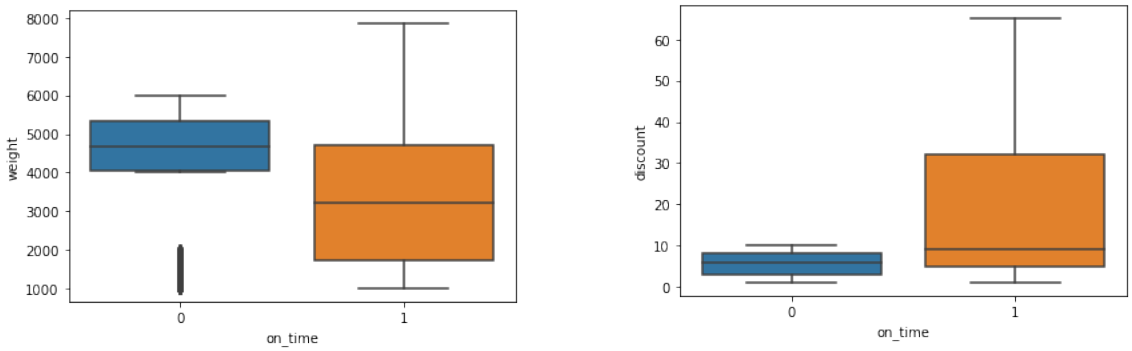

Fig 7

To summarize, this analysis has found that the variables that have more influence when defining the success of the logistics system are “weight” and “discount”.

References

Chen, C., Xu, X., Zou, B., Peng, H., & Li, Z. (2021). Optimal decision of multiobjective and multiperiod anticipatory shipping under uncertain demand: A data-driven framework. *Computers & Industrial Engineering*, *159*, 107445. https://doi.org/10.1016/j.cie.2021.107445

EMC Education Services. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data* (1st ed.). Wiley.

Ponce, D., Contreras, I., & Laporte, G. (2020). E-commerce shipping through a third-party supply chain. *Transportation Research Part E: Logistics and Transportation Review*, *140*, 101970. https://doi.org/10.1016/j.tre.2020.101970

Appendix

Dataset extracted from: https://www.kaggle.com/prachi13/customer-analytics

Fig1, Fig 2, Fig 3, Fig 4, Fig 5, Fig 6, Fig 7 were extracted from the analysis on Jupyter Notebook in Python. Find the code attached to this report.