Data Wrangling

- ONLINE PURCHASE INTENTION -

By: Tanvi Gandhi

MOTIVATION

Possible questions to answer during Data Wrangling:

- What kind of cleaning steps were performed?
- How the missing values were dealt with, if any?
- Were there outliers, and how were they handled?

DATA INFORMATION

The data has 12330 rows and 18 feature vectors including the categorical target feature. Among them there are 10 numerical and 8 categorical features.

Below is a brief explanation of the variables: -

Numerical features

Administrative: Total number of administration related pages visited by the visitor in that session like account management.

Informational: Total number of Informational pages visited by the visitor in that session like information about offers, cart etc.

ProductRelated: Total number of product related pages visited by the visitor in that session.

Note: The values of the above features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

Administrative_Duration: Total amount of time (in seconds) spent by the visitor on adminitration related pages.

Informational_Duration: Total amount of time (in seconds) spent by the visitor on information related pages.

ProductRelated_Duration: Total amount of time (in seconds) spent by the visitor on product related pages.

Note: The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.

BounceRates: Average bounce rate value of the pages visited by the visitor. The value of Bounce Rate feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.

ExitRates: Average exit rate value of the pages visited by the visitor. The value of Exit Rate feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session.

PageValues: Average page value of the pages visited by the visitor. The Page Value feature represents the average value for a web page that a user visited before landing on the goal page or completing an Ecommerce transaction (or both).

SpecialDay: Closeness of the site visiting time to a special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

Categorical features

Month: Month value of the visit date

OperatingSystems: Operating system of the visitor

Browser: Browser of the visitor

Region: Geographic region from which the session has been started by the visitor

TrafficType: Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)

VisitorType: Visitor type as 'New Visitor', 'Returning Visitor' and 'Other'

Weekend: Boolean value indicating whether the date of the visit is weekend

Revenue: Class label indicating whether the visit has been finalized with a transaction

DATA CLEANING

The data had some inconsistency and unavailable data. Both of these errors could not be handled. Below are the explanations:

- The data is available for 10 months (i.e. Jan and April not included). This discrepancy cannot be handled.

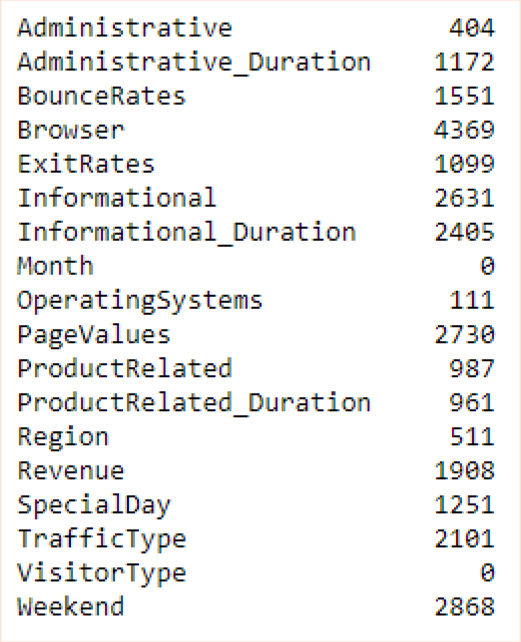
- There seems to be some sort of inconsistency in the special days data since it does not have any values for December month - which has Christmas and New Year as the biggest holidays. This cannot be imputed since no data is available.
- The target vector (binary True and False) is highly imbalanced as the False values are as much as 5 times of True Values. To handle this, we will perform SMOTE analysis in the pre-processing phase.
- The data contains 125 duplicate rows. However, we can keep them since it is only around 1% of the data and isn't necessarily erroneous. Further, we also need to realize the reasons due to which the data can seem duplicate:
  1. It is possible that two different sessions of two different users can be same i.e. they visited the exact same pages and may have same browser or OS type and reside in the same region. This is rare event but not impossible.
  2. It is likely that the same user (or a relative) could perform the same set of page flow from other device with same configuration of OS and browser at around the same time.

MISSING VALUES

Surprisingly, there were no missing values in the dataset.

OUTLIERS

Below is a screenshot of detected outliers using IQR method:



```
Administrative              404
Administrative_Duration    1172
BounceRates                1551
Browser                    4369
ExitRates                  1099
Informational              2631
Informational_Duration     2405
Month                         0
OperatingSystems            111
PageValues                 2730
ProductRelated              987
ProductRelated_Duration     961
Region                      511
Revenue                    1908
SpecialDay                 1251
TrafficType                2101
VisitorType                   0
Weekend                    2868
```

We can exclude the categorical features from the outlier’s consideration since it doesn't make sense to have outliers for them. For e.g., Weekend column might detect outliers since there are too many False values as there are True values. But this scenario is natural in real life as there are a greater number of non-weekends days than weekend days.

Coming to the categorical data, the high numbers of outliers suggest that they might be natural outliers (as opposed to measurement error) so it would be okay to keep them. For e.g. outliers for BounceRates are 1551. It's not surprising to think that a frequently bounced page might sometime yield revenue. Also, a page with low bounce rate might not always result in revenue. So, these measurements might be detected as outliers but are normal if considered in real-life.

In our case, we will do nothing with outliers since all the numerical features are metrics and are suggestive of natural outliers. Thus, removing them can result in loosing important information. We can later on apply Standardization to bring the values to the same scale.