Predicting Online Shoppers Purchasing Intention

# Introduction

In recent years, one of the most popular and rapidly expanding buying methods throughout the world has been online shopping. This is demonstrated in the growing percentage of customers who shop online as well as recent growth in online retail sales. Despite this, the percentage of consumers who make purchases online and then immediately leave the website is still far greater than what e-commerce platforms had predicted it would be (Sakar et al., 2018)[C O19]. For this reason, as the internet and e-commerce continue to expand at a rapid pace, it is imperative that online merchants anticipate the aspects that impact customer intent to buy online.

In addition, the growth of the internet and e-commerce has had an effect on the lives of users, the manner in which they traded, and the process by which they made decisions, which has led to the creation of a distinction between the behavior of consumers who engage in online consumption and those who engage in behavior associated with conventional consumption. At the same time, it has become increasingly important for online retailers to be familiar with the behaviors and objectives of the many sorts of clients they serve (Kim Kim, 2004)[Kim04].Analyzing the past transactions of clients allows one to make educated guesses about their future purchasing behavior. Since of this, and because we wanted to understand more about the elements that influence the buying behavior and intention of consumers, we decided to undertake particular research using the dataset titled ”Online Shoppers Purchasing Intention Dataset.”

This study aims to provide a global picture of the numerous components that a platform might employ to enhance its customer decision-making process. It demonstrates how the various components of the platform may be employed to increase the efficacy of the platform. Several different categorization models for predicting consumer purchase intention have been developed with the use of the data from online shopping activity. The prediction model that was developed may be utilized in a variety of facets of a website, such as the prediction of a user’s upcoming intentions to make a purchase.

Data science gives businesses the ability to monitor, manage, and record performance measures to support better decision-making throughout the entire enterprise. Trend analysis enables businesses to take important decisions that will improve consumer engagement, raise productivity levels, and boost profits. So, it enables us to combine all of our knowledge and showcase that through a major project.

# Problem statement

E-commerce websites are responsible for around 9 percent of all retail sales that take place in the United States. In point of fact, businesses like Amazon have established retail empires as a direct result of operating such massive online marketplaces. It is vital for businesses that operate in e-commerce industry to have a good awareness of the dynamics that impact the client purchase intention in order to be successful given the competition of the e-commerce platforms and their demand today.

In addition, the businesses should be able to influence those dynamics in their favour to increase the likelihood that potential customers go through with the transactions. Exploring the online historic purchase data may lead to the discovery of critical information that, in turn, may lead to increased sales by influencing customer purchase intent. The potential of e-commerce to sway the intent of customers to make a purchase is currently hidden in the data. This is one of the reasons we felt it would be essential to investigate the ’Online Shoppers Purchasing Intention Dataset’ to gain the insight and predict the purchasing behaviour of customers on a particular website.

# Objectives

The objectives of the study include:

* To determine the elements that influence consumer purchase intention most.
* To Understand the consumer purchasing behavior.
* To Develop a model for predicting the consumer intent to purchase

# Research design and methodology

We are going to investigate the ”Online Shoppers Purchasing Intention Dataset” using python for data science in order to gain insight and information that is essential for decision making. The modelling procedure consisted of the following seven steps:

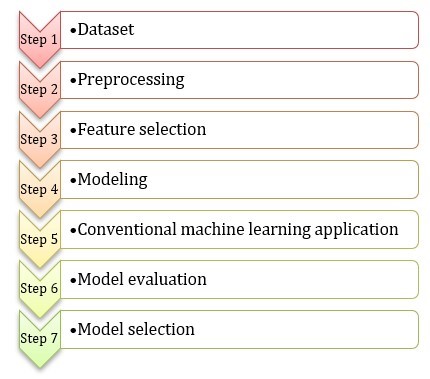


Figure 1: Modelling Procedure

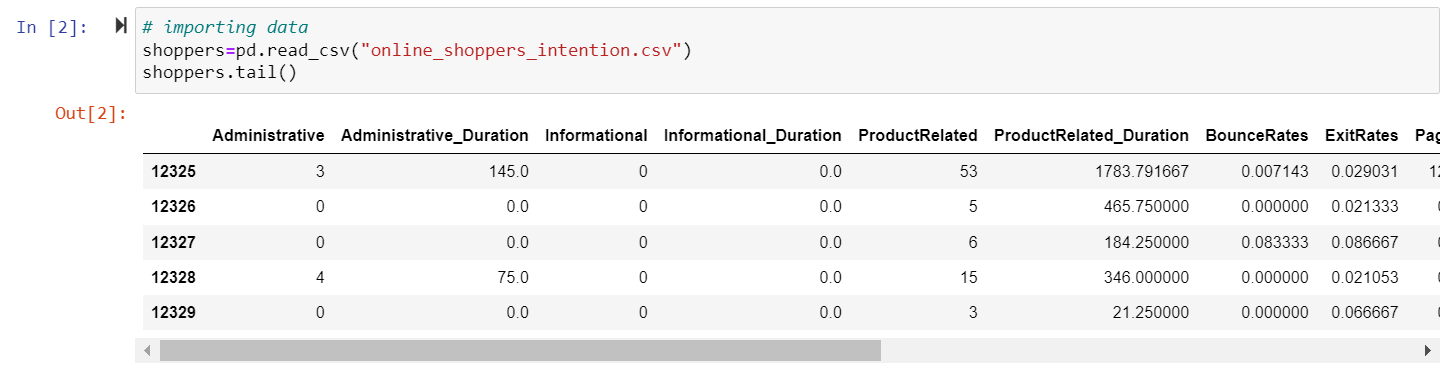
First, the data Set is cleaned, and then the categorical variables are coded such that they are compatible with the various classification algorithms. The users are categorized according to the likelihood that they will create revenue, and then various Machine learning algorithms, such as Support Vector Machine (SVM), logistic classifier, as well as Voting Classifier, are used to forecast whether or not the users would make a purchase. Methods such as random oversampling and SMOTE are utilized in addressing concerns with the imbalance of classes. In addition, after experimenting with both classical algorithms such as tree-based algorithms and SVM, and deep learning techniques such as multilayer perceptron, we chose the method that ranked highest to predict the purchasing intention.

# Dataset

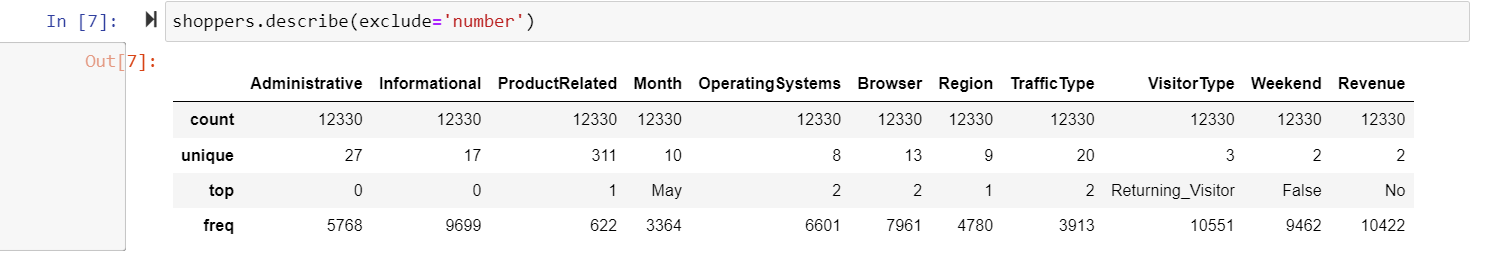
The data was sourced from UCI machine learning repository whose link is attached in the appendices. The dataset contains 12,330 observations, each of which represents a visit to a website for the purpose of online buying. The data span a period of one year to exclude the possibility of any trend. Each observation is characterized by a total of 18 attributes, which are then subdivided into ten numerical and 8 category aspects. The name of our binary classification feature is ”Revenue,”. Our objective is to make use of the other 17 features listed below in order to make a prediction about the label ”Revenue,” which means to determine whether or not a visit session will result in a transaction.

EDA analysis

The ‘online\_shoppers\_intention’ dataset was successfully selected for analysis using read csv module in pandas package as shown in the code below.



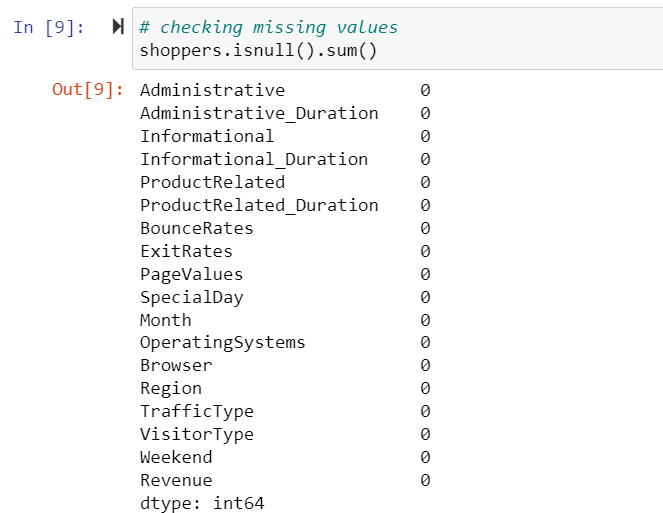
Data description for non-numeric data



From the figure above, it is evident that 11 variables were non numeric. The feature of most importance is revenue which is a binary data type and the dependent variable in this analysis. ‘No’ is the most frequent value in revenue feature with a count of 10422 out of 12330 (84.5%). Such variation between binary variations are likely to cause biasness due to class imbalance.

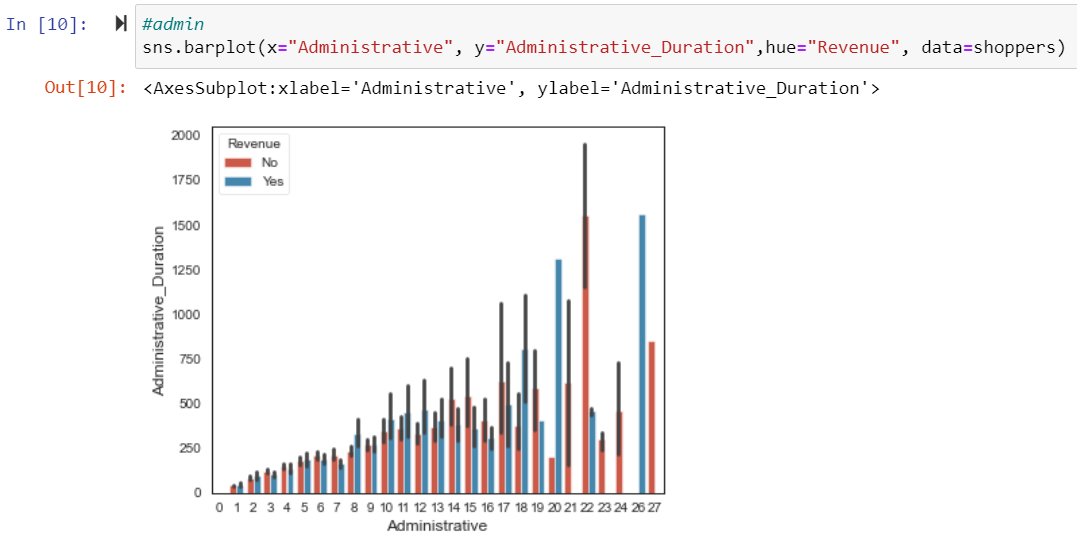
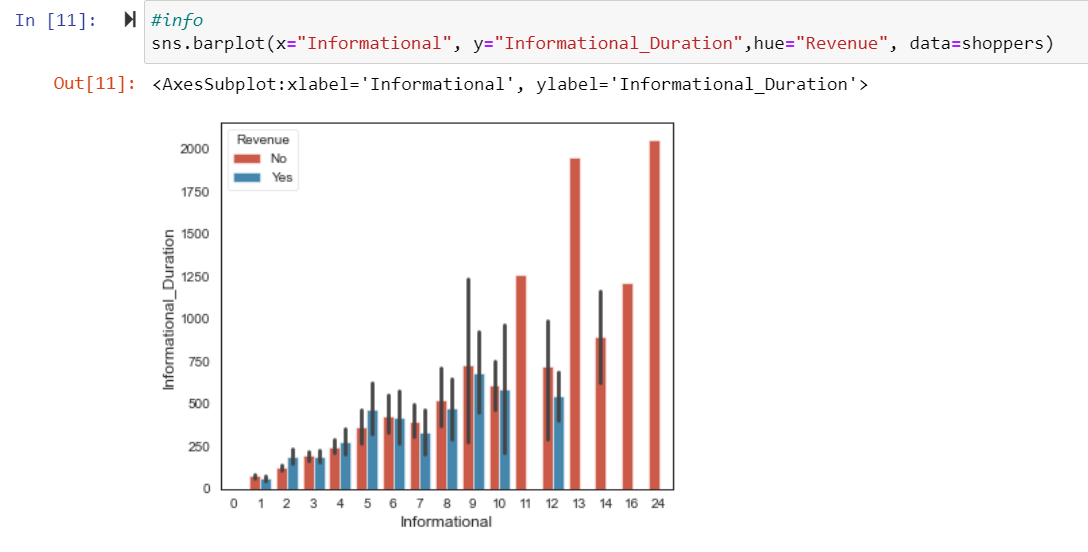
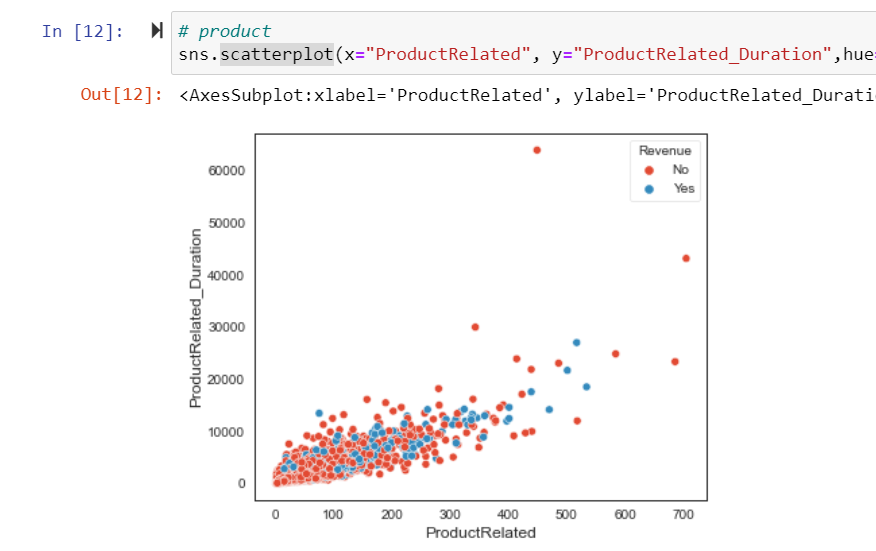
Missing values

In nearly all cases of research, there are gaps in the data, even when the study is well-conducted and monitored. The statistical power of an investigation can be diminished by missing observations in a dataset, which can also lead to erroneous estimates and, ultimately, inaccurate findings. This book discusses the issues that might arise from missing value, the many forms of missing data, as well as the methods for dealing with missing observations. Nevertheless our dataset had no missing values.



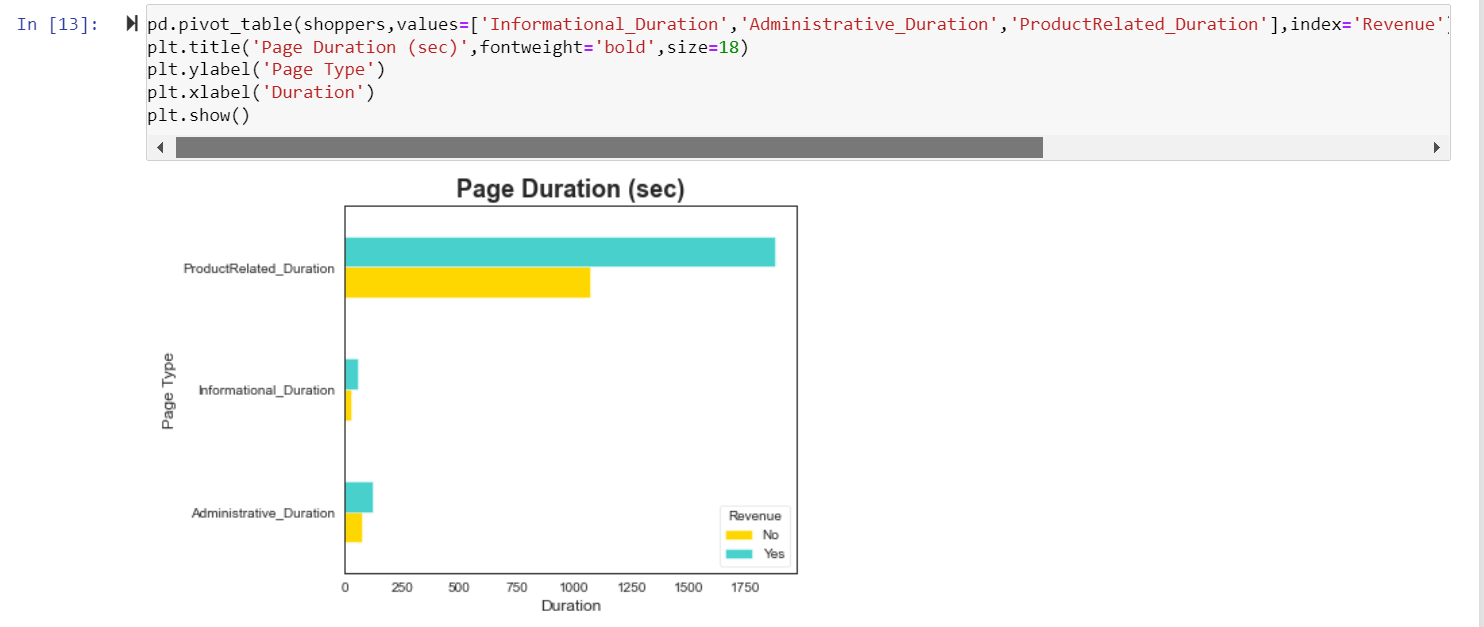
Visualization

Duration on site

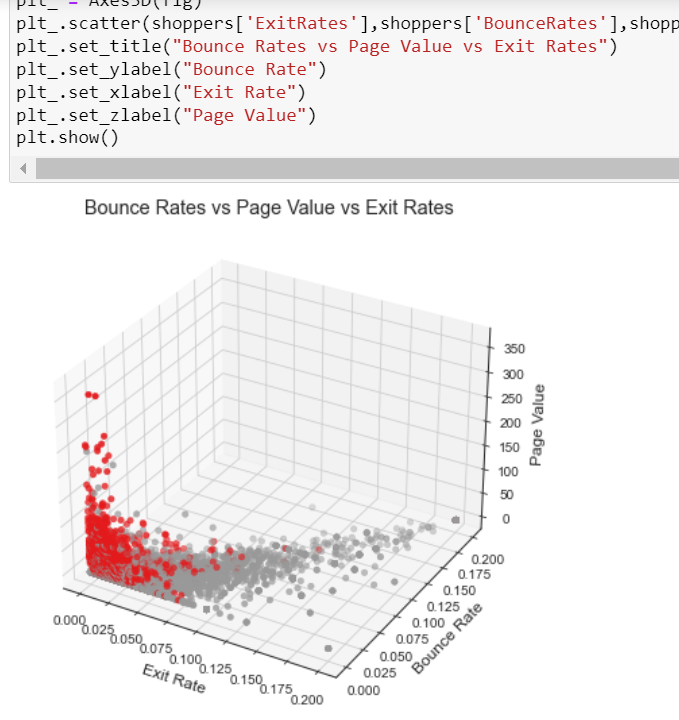
From the plots above, it is evident that there is no multicollinearlity between product related and product duration. In addition, the pages relating to products were visited throughout the vast majority of people's time spent on the website. Additionally, it provides the greatest contribution to the overall generating of income. According to the data provided by the number of visits made by customers, product pages that are directly connected to the product are of the biggest significance to the customer.

Duration vs Revenue



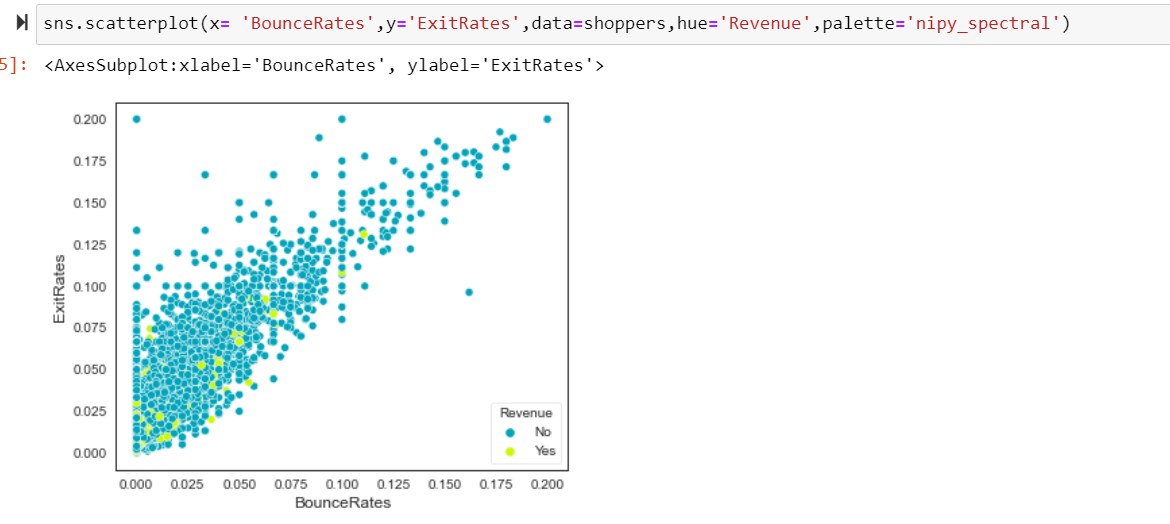
The pages related to products received the vast majority of visitors' attention while they were on the website. As a consequence of this, product-related pages are of the biggest significance to the customer from the point of view of the generation of income.

Bounce, Exit Rate vs Page value

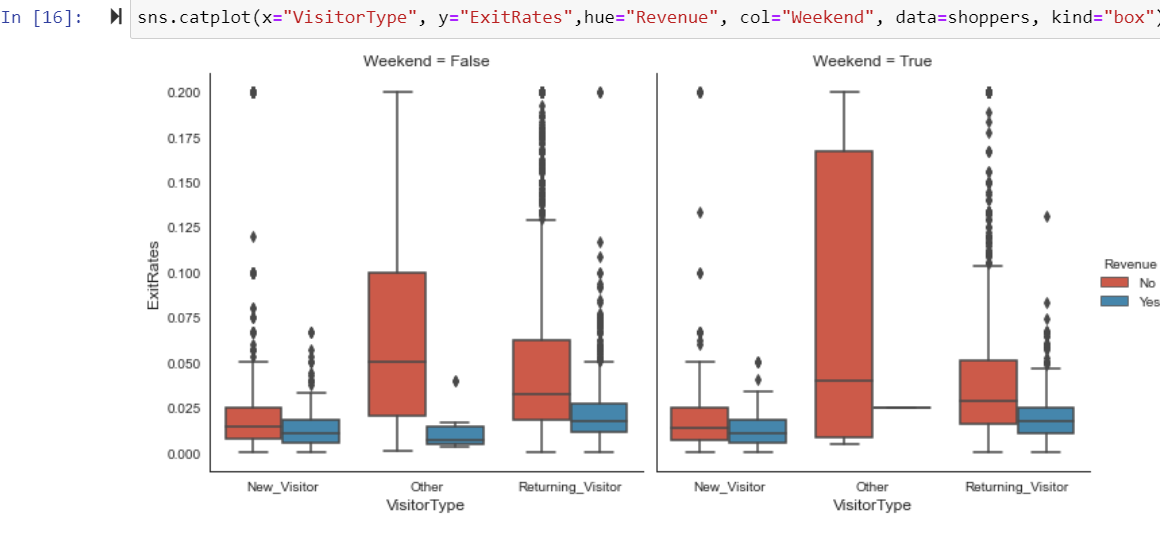


Customers who did not contribute to the company's revenue are represented by red points, while customers who did contribute to the company's revenue are represented by black points. Additionally, compared to black points, clients who actually made a purchase (identified to with reds points) had a bounce rate and exit rate that are far lower. Those who ended up making purchases have a page value that is significantly higher compared to customers who did not make purchases.

Is there relation between Exit rate and Bounce Rate?



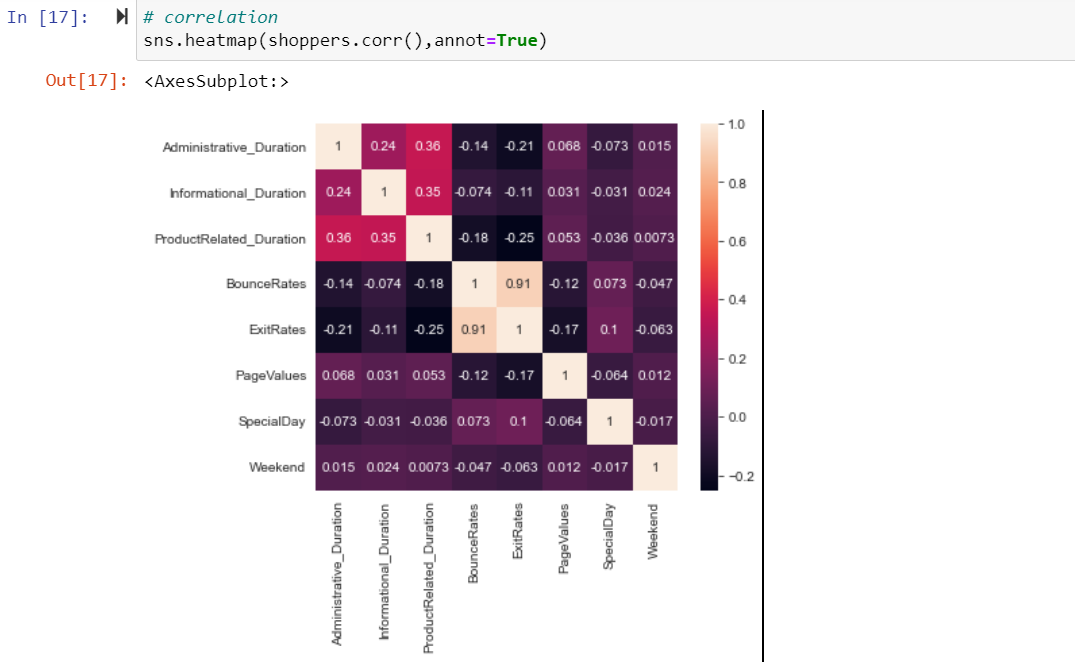
Exit rate and bounce rate had a positive correlation. Meaning that high bounce rate is associated with high exit rate.



When there is a revenue, the exit rates have a very low spread, and there isn't much of a difference in exit rates when you take into account the types of visitors as well as the weekend. When it's the weekend and there is no revenue, there is a significant difference in the exit rates that are found in other categories. It's possible that they're more of the window shopping kind. New visitors have low leave rates, which are very consistent regardless of the revenue scenario. It works quite well to keep the new customers coming back.

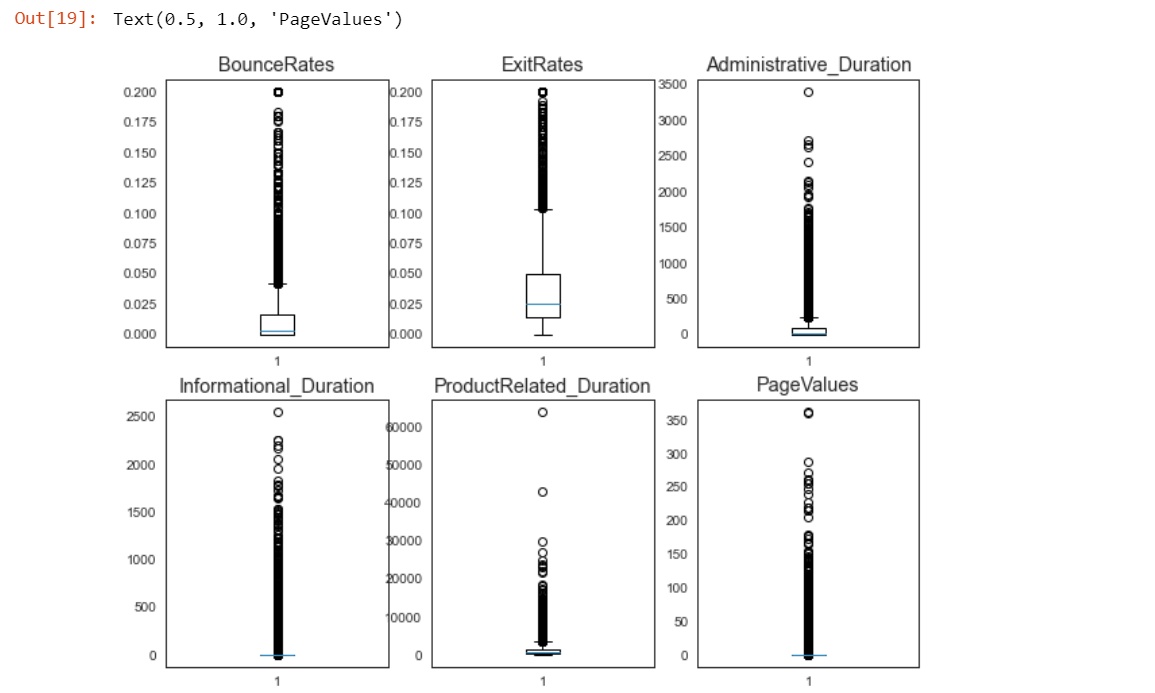
**Data analytics**

**Multicollinearlity check**



There is a strong association between the values of each page and the amount of revenue made. Pages that have high bounce levels also generally have higher exit levels, which is another factor that has a negative impact on revenues. Pages on the website that are related to products earn a substantial amount of revenue. Hence, we have to drop one of them in the modelling process to avoid multicollinearlity.

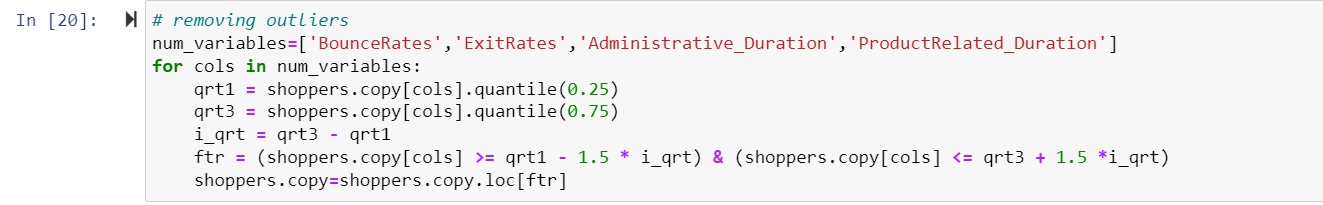
Checking for outliers



It is patently obvious that there are a great deal of outliers. Informational duration and page values do not contain any outliers, and if you delete the values that are considered to be outliers, there will only be one value remaining in the set. Therefore, with the exception of those two characteristics, we will be eliminating the outliers using the IQR approach.

**Removing outliers**

Presence of outliers has a multiplicative effect on error variance as well as dilution effect on the power of analytical techniques. They have the potential to introduce bias and/or to alter estimates. They can also have an effect on the fundamental assumption of extrapolation, in addition to having an effect on other statistical models. Outliers present in the data sample are removed using the code below:



Data scaling

Hot encoding

Feature selection

**Data visualization and results report**

Online learning algorithms can be described as passive-aggressive algorithms. In the case that the classification is performed correctly, this kind of algorithm does not make any changes or updates; nevertheless, it becomes more active in the event that the computation is incorrect. This algorithm, in contrast to the vast majority of others, does not converge. Its goal is to apply updates that will reverse the loss while having very little impact on the norm of the weight vector. The Passive Aggressive algorithm works wonderfully when it comes to categorizing vast amounts of data streams. It is simple to create and extremely quick, but it does not give global guarantees in the same way that the support-vector machine does (SVM). When applied to our data, the Passive Aggressive Classifier achieved an accuracy rating of 84.21 percent both on the test set and on the validation set.

Support Vector Machines (SVMs) are a supervised learning approach generally used for classification, and also suited for certain regression problems. The fundamental SVM method is a binary linear classifier. This classifier places previously unseen data points into one of two groups based on a set of labelled "training" points, and it does so by creating a linear split between the two groups. The Support Vector Classifier achieved a slightly greater level of accuracy on both the test set and the validation set than the Passive Aggressive Classifier did, scoring 88.51 percent on both.

For random forest as the name suggest, the model is comprised of a number of decision trees which operate in harmony as an ensemble where each tree forecasts a given class. The class that receives the most votes is the one that serves as the basis for our model's prediction. On both the test set and the validation set, the Random Forest Classifier achieved an accuracy score of 90.94 percent, which was higher than the scores achieved by either of the two earlier models (Passive Aggressive and Support Vector Classifiers).

# Conclusion

Given the problem statement above, it is clear that accurately forecasting the purchase behaviour and intentions of customers is of the utmost importance. The provision of these insights through the acquisition of historical and current data can assist e-commerce platforms in lowering their bounce rate and increasing the number of customers who make purchases on their sites. This allows e-commerce enterprises to increase their revenue by appropriately addressing the factors that influence their customers’ purchase intents.

# References

|  |  |
| --- | --- |
| [C O19] | Mete Katircioglu Yomi Kastro C. Okan Sakar S. Olcay Polat. “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks”. In: *Neural Computing and Applications* 31 (2019), pp. 6893–6908. |
| [Kim04] | Young Kim E Kim Y. “Predicting online purchase intentions for clothing products”. In: *European Journal of Marketing* 38.7 (2004), pp. 883–897. |