Predicting Online Shoppers Purchasing Intention

*

Pavanika Thotakura - 11441164
Department of Information Science
University of North Texas
Denton, TX, US
pavanikathotakura@my.unt.edu

Vaishnavi Baireddy - 11450357
Department of Information Science
University of North Texas
Denton, TX, US
vaishnavibaireddy@my.unt.edu

Jalendar Reddy Maligireddy - 11511290
Department of Information Science
University of North Texas
Denton, TX, US
jalendarreddymaligireddy@my.unt.edu

Abhishek Reddy Boddu - 11442213
Department of Information Science
University of North Texas
Denton, TX, US
abhishekreddyboddu@my.unt.edu

I. Introduction and Statement of the Problem

A. 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) [1]. 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 behaviour of consumers who engage in online consumption and those who engage in behaviour associated with conventional consumption. At the same time, it has become increasingly important for online retailers to be familiar with the behaviours and objectives of the many sorts of clients they serve (Kim Kim, 2004) [2].Analysing the past transactions of clients allows one to make educated guesses about their future purchasing behaviour. Since of this, and because we wanted to understand more about the elements that influence the buying behaviour 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.

B. 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.

II. Review of Literature

In this research paper, authors provided a real-time online shopper behavioural assessment system made up of two modules that predicts both the likelihood of website desertion and the visitor's purchase intent. To increase the efficiency and scalability of the classifiers, they pre-processed the data using feature selection and oversampling. The findings demonstrate that compared to RF and SVM, MLP, which is calculated using a resilient backpropagation algorithm with weight backtracking, produces much greater accuracy and F1 Score. The likelihood estimate of a visitor's intention to depart the site before completing the transaction is shown in a sigmoid output in the second module utilizing only sequential clickstream data that was used to train the long short-term memory-based recurrent neural network.

Their results provide credence to the claim that attributes gleaned from visit clickstream data provide crucial information for predicting online purchase intention. The filter feature ranking algorithms place the features that represent aggregated statistics of the clickstream data acquired during the visit close to the top.

In order to account for these redundancies between the features, they used a feature ranking method termed minimum redundancy-maximum relevance. The results demonstrated that selecting a basic subset of aggregated statistics from clickstream data and session data, such as the date and location, leads to a more precise and scalable solution. The second module of the suggested system trains an LSTMRNN to forecast the likelihood that a user would depart the website within the specified prediction horizon using just sequential clickstream data. The amount of pageviews a visitor will complete before leaving a website determines the prediction horizon. A key finding is that when the prediction horizon gets longer, it takes longer and requires more steps for the user to decide to leave, making estimate harder and the success rate decreasing.

Online purchasing for apparel is becoming more and more common. Based on online shopping characteristics and demographic data, this research paper established dimensions of online shopping characteristics and predictors of intention to purchase clothing, jewellery, or accessories. The four criteria that made up the four perceived qualities of online buying were the transaction/cost, incentive programs, site design, and interaction. The association between education level and intention to make an online purchase was also mediated by the incentive scheme.

The mean scores for each of the elements that made up this factor were all quite high. Additionally, compared to the other three components, this factor's explained variation was by far the largest. This suggests that the transaction/cost component is the main driver of how people perceive the various aspects of online shopping

III. Objectives of the Study

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

IV. 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 modeling procedure consisted of the following seven steps:

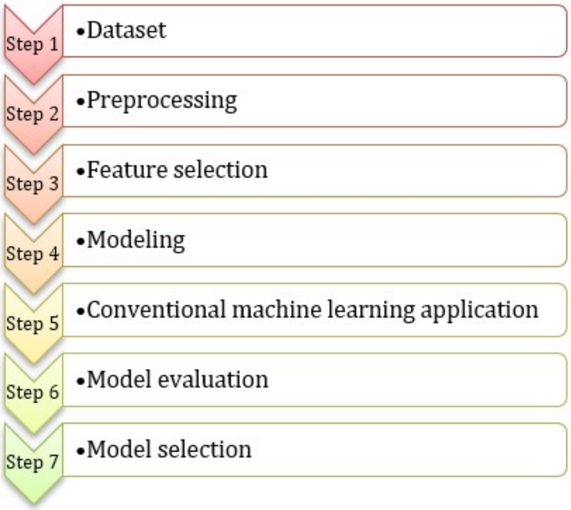


Fig. 1. Modeling Procedure

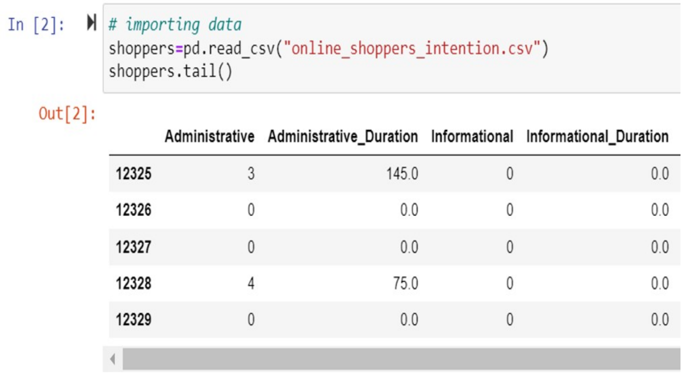
First, the data Set is cleaned, and then the categorical variables are coded such that they are compatible with the various classification algorithms. The users are categorized according to the likelihood that they will create revenue, and then various Machine learning algorithms, such as Support Vector Machine (SVM), passive - aggressive classifier and Random forest Classifier, are used to forecast whether or not the users would make a purchase. In addition, after experimenting with both classical algorithms such as tree-based algorithms and SVM, and online-learning algorithms such as Passive-Aggressive classifier, we chose the method that ranked highest to predict the purchasing intention.

V. Data Collection

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 dataset contain one year of data so that we may not lose 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. url: https://www.kaggle.com/datasets/henrysue/online-shoppers-intention

VI. Exploratory data analysis (EDA)

The ‘online_shoppers_intention’ dataset was successfully selected for analysis using read_csv module in pandas package as shown in the code below.

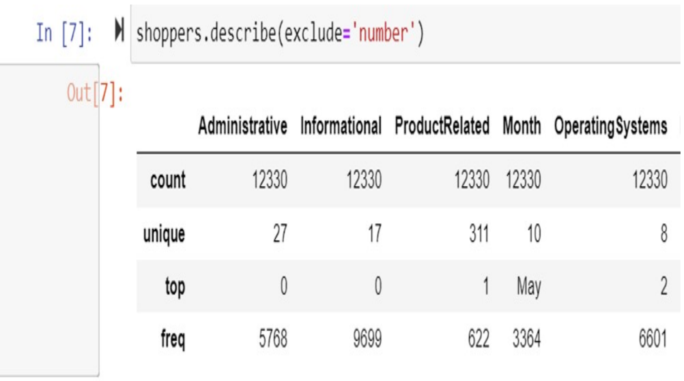


```
In [2]:  # importing data
         shoppers=pd.read_csv("online_shoppers_intention.csv")
         shoppers.tail()
Out[2]:
```

	Administrative	Administrative_Duration	Informational	Informational_Duration
12325	3	145.0	0	0.0
12326	0	0.0	0	0.0
12327	0	0.0	0	0.0
12328	4	75.0	0	0.0
12329	0	0.0	0	0.0

Fig. 2. Data Description

Data description for non-numeric data:



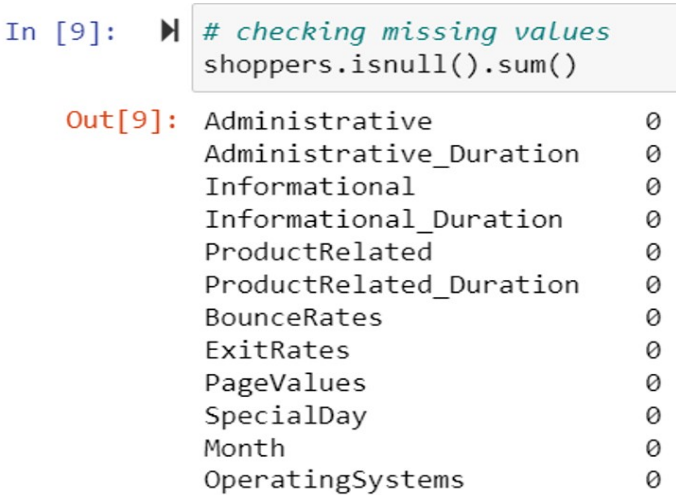
```
In [7]:  shoppers.describe(exclude='number')
Out[7]:
```

	Administrative	Informational	ProductRelated	Month	OperatingSystems
count	12330	12330	12330	12330	12330
unique	27	17	311	10	8
top	0	0	1	May	2
freq	5768	9699	622	3364	6601

Fig. 3. 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.

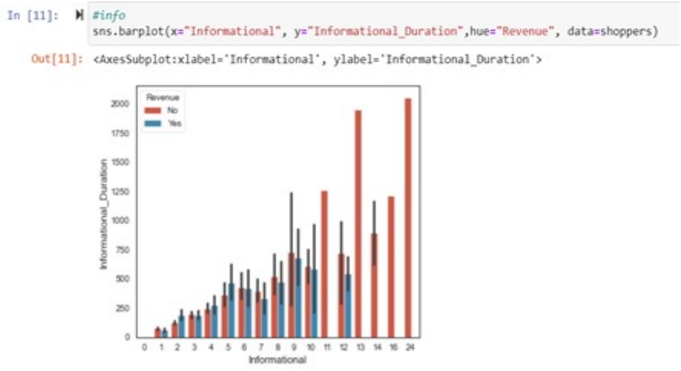
Missing values : In nearly all cases of research, there are gaps in the data, even when the study is well-conducted and monitored. The statistical power of an investigation can be diminished by missing observations in a dataset, which can also lead to erroneous estimates and, ultimately, inaccurate findings. 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.



```
In [9]:  # checking missing values
         shoppers.isnull().sum()

Out[9]:  Administrative              0
         Administrative_Duration     0
         Informational               0
         Informational_Duration      0
         ProductRelated              0
         ProductRelated_Duration     0
         BounceRates                 0
         ExitRates                   0
         PageValues                  0
         SpecialDay                  0
         Month                       0
         OperatingSystems            0
```

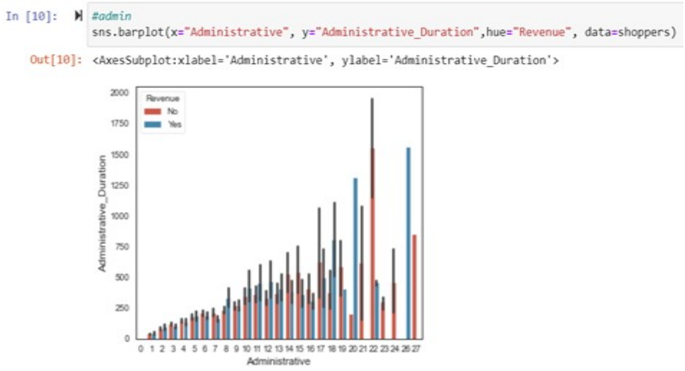
Fig. 4. No Missing Values in the Dataset



```
In [11]:  #info
          sns.barplot(x="Informational", y="Informational_Duration",hue="Revenue", data=shoppers)

Out[11]: <AxesSubplot:xlabel='Informational', ylabel='Informational_Duration'>
```

Fig. 5. Duration on site

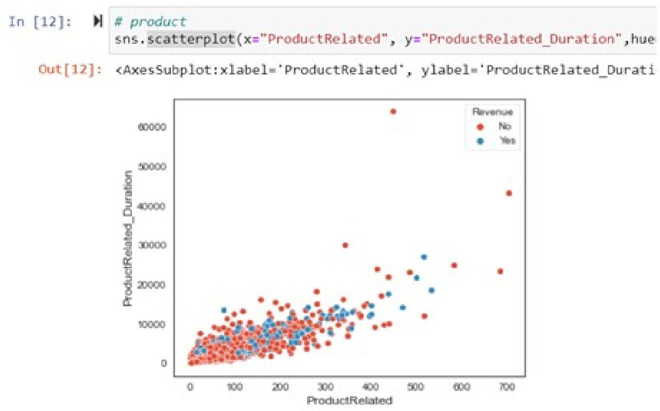


```
In [10]:  #admin
          sns.barplot(x="Administrative", y="Administrative_Duration",hue="Revenue", data=shoppers)

Out[10]: <AxesSubplot:xlabel='Administrative', ylabel='Administrative_Duration'>
```

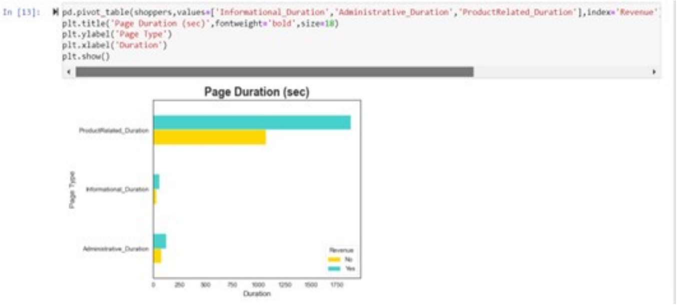
Fig. 6. Duration on site

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.



```
In [12]:  # product
          sns.scatterplot(x="ProductRelated", y="ProductRelated_Duration",hue
Out[12]:  <AxesSubplot:xlabel='ProductRelated', ylabel='ProductRelated_Durati
```

Fig. 7. Duration on site



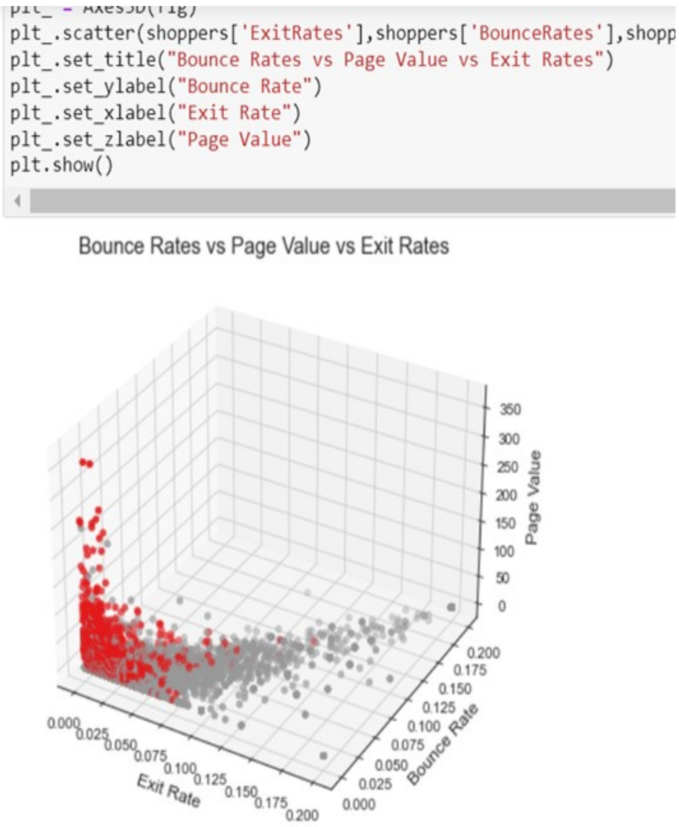
```
In [13]:  pd.pivot_table(shoppers,values=['Informational_Duration','Administrative_Duration','ProductRelated_Duration'],index='Revenue'
          plt.title('Page Duration (sec)',fontweight='bold',size=18)
          plt.ylabel('Page Type')
          plt.xlabel('Duration')
          plt.show()
```

Fig. 8. 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.

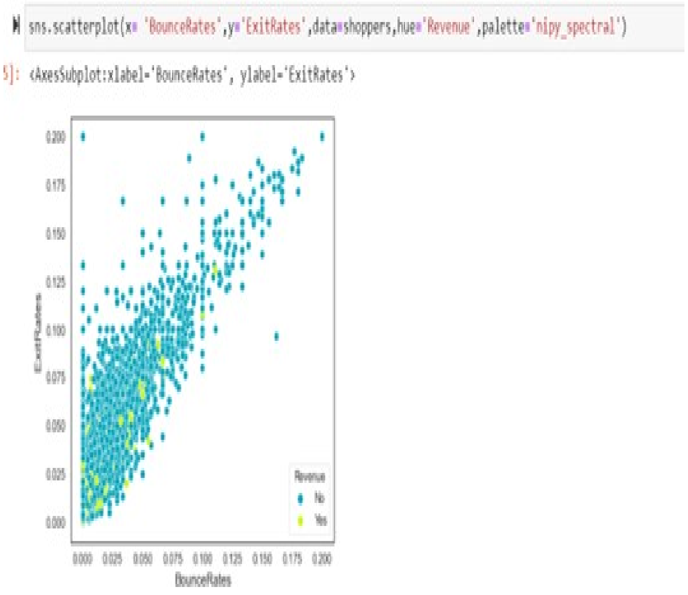
Customers who did not contribute to the company's revenue are represented by red points, while customers who did contribute to the company's revenue are represented by black points. Additionally, compared to black points, clients who actually made a purchase (identified to with reds points) had a bounce 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.

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



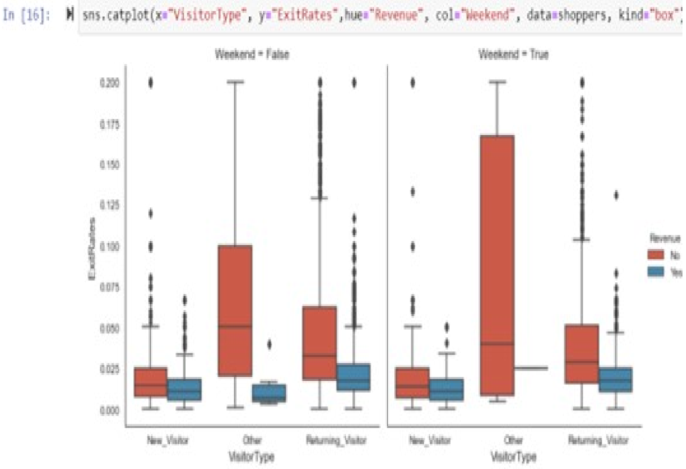
```
plt_ = Axes3D(fig)
plt_.scatter(shoppers['ExitRates'],shoppers['BounceRates'],shopp
plt_.set_title("Bounce Rates vs Page Value vs Exit Rates")
plt_.set_ylabel("Bounce Rate")
plt_.set_xlabel("Exit Rate")
plt_.set_zlabel("Page Value")
plt.show()
```

Fig. 9. Bounce, Exit Rate vs Page value



```
sns.scatterplot(x= 'BounceRates',y='ExitRates',data=shoppers,hue='Revenue',palette='nipy_spectral')
5]: <AxesSubplot:xlabel='BounceRates', ylabel='ExitRates'>
```

Fig. 10. Is there relation between Exit rate and Bounce Rate?

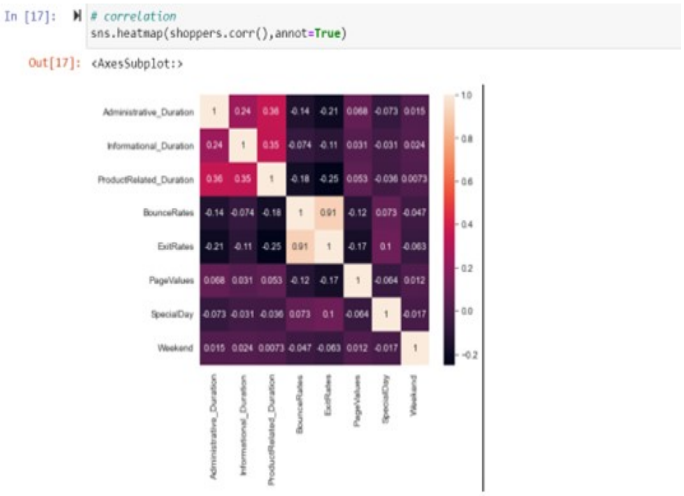


```
In [16]:  sns.catplot(x="VisitorType", y="ExitRates",hue="Revenue", col="Weekend", data=shoppers, kind="box",
```

Fig. 11.

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.

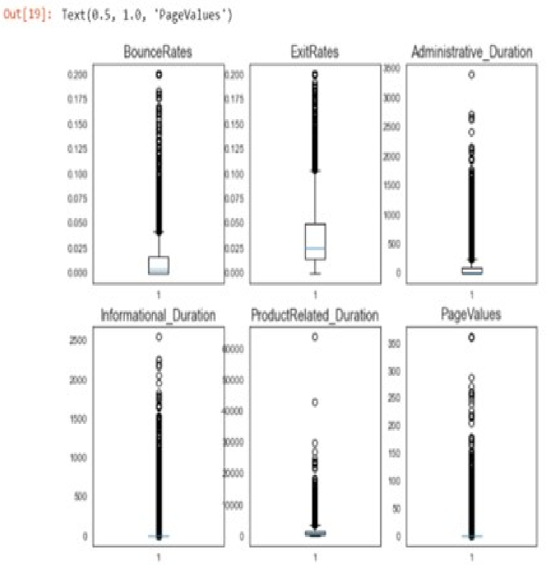
VII. Data Analytics / Modeling



```
In [17]:  # correlation
          sns.heatmap(shoppers.corr(),annot=True)
Out[17]: <AxesSubplot:>
```

Fig. 12. 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.

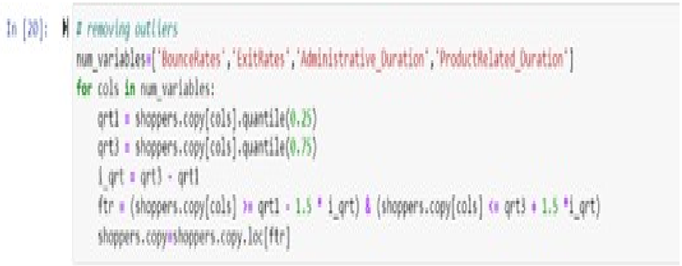


```
Out[19]: Text(0.5, 1.0, 'PageValues')
```

Fig. 13. 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.

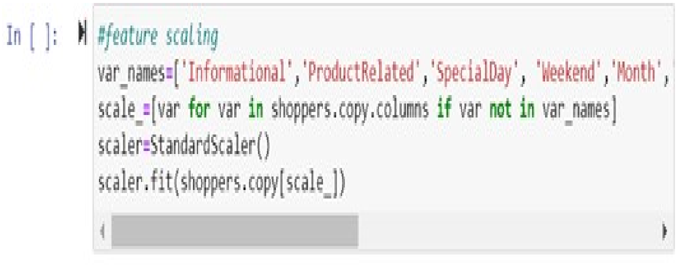
Presence of outliers has a multiplicative effect on error variance as well as dilution effect on the power of analytical techniques. They have the potential to introduce bias and/or to alter estimates. They can also have an effect on the fundamental assumption of extrapolation, in addition to having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



```
In [20]:  # removing outliers
          num_variables=['BounceRates','ExitRates','Administrative_Duration','ProductRelated_Duration']
          for cols in num_variables:
              qrt1 = shoppers.copy[cols].quantile(0.25)
              qrt3 = shoppers.copy[cols].quantile(0.75)
              i_qrt = qrt3 - qrt1
              ftr = (shoppers.copy[cols] >= qrt1 - 1.5 * i_qrt) & (shoppers.copy[cols] <= qrt3 + 1.5 *i_qrt)
              shoppers.copy=shoppers.copy.loc[ftr]
```

Fig. 14. Removing outliers

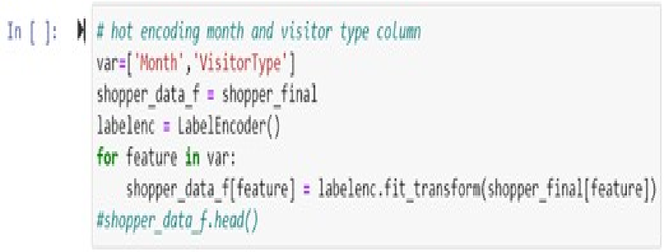
When we are applying machine learning algorithms to a data set, one of the procedures that falls under the umbrella of "data pre-processing" is scaling the data. Many of machine learning techniques draw conclusions based on the data sets they're given, and programs commonly measure the distance between observations especially in classification to make better predictions. When performing regression modelling, scaling the dependent variable is a smart concept; scaling the data makes it simple for a model to learn as well as comprehend the issue at hand. In the context of neural networks, the presence of an independent variable whose values are distributed throughout a range can lead to a significant loss both during training and testing and make the learning process unstable. The machine learning models assign different weights to the input variables based on the data points and inferences that are produced by the output. If this is the case, and the disparity in value between the data points is significant, the model will have to assign a greater weight to each of the points; nevertheless, when the results are tallied, a model with a high weight value variation frequently exhibits instability. This indicates that the model could give inaccurate results or might not perform well throughout the learning process. In this analysis, standardscaler() is used as shown below:



```
In [ ]:  #feature scaling
         var_names=['Informational','ProductRelated','SpecialDay', 'Weekend','Month',
         scale_=[var for var in shoppers.copy.columns if var not in var_names]
         scaler=StandardScaler()
         scaler.fit(shoppers.copy[scale_])
```

Fig. 15. Data scaling

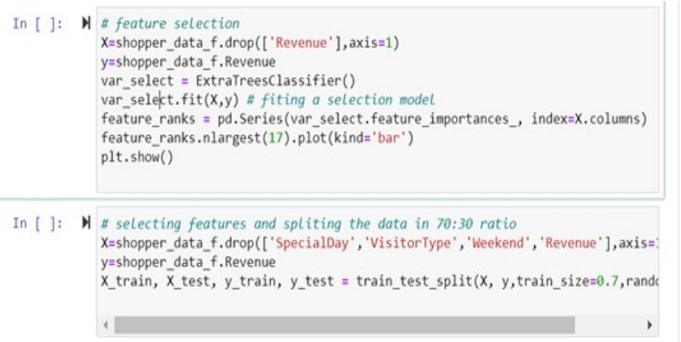
When categorical data attributes are converted to numbers using one hot encoding, it improves predictive performance as well as classification precision. One Hot Encoding is a method that is frequently used for pre-processing categorical features in preparation for machine learning models. When categorical data is introduced directly into a model, a number of machine learning algorithms, including Deep Learning Algorithms, are typically unable to function properly. These categories need to be further turned into numbers, just like the categorical input and output variables in the data need to go through the same process. In this analysis one Hot encoding is used as shown below:



```
In [ ]:  # hot encoding month and visitor type column
         var=['Month','VisitorType']
         shopper_data_f = shopper_final
         labelenc = LabelEncoder()
         for feature in var:
             shopper_data_f[feature] = labelenc.fit_transform(shopper_final[feature])
         #shopper_data_f.head()
```

Fig. 16. Hot encoding

High-dimensional data analysis presents scientists and engineers working in the disciplines of data mining and machine learning (ML) with a significant obstacle to overcome. Reducing redundant as well as unnecessary data may be accomplished through the use of a feature selection, which is a technique that is equally basic and efficient. By eliminating redundant and unwanted data boosts the learning accuracy and decreases the code runtime which also helps in understanding the results better. When creating an ML model in the real world, it is common practice to ignore some of the variables in the dataset since they are not valuable. When redundant variables are added to a model, the generalization competency of the model suffers, and there is a possibility that the overall precision of a classifier is also reduced. Additionally, the development of a more complicated model follows the addition of additional variables to an existing one. In this analysis we used Extra Trees Classifier model to find the ranking of feature importance and we selected the 13 best features.



```
In [ ]:  # feature selection
         X=shopper_data_f.drop(['Revenue'],axis=1)
         y=shopper_data_f.Revenue
         var_select = ExtraTreesClassifier()
         var_select.fit(X,y) # fiting a selection model
         feature_ranks = pd.Series(var_select.feature_importances_, index=X.columns)
         feature_ranks.nlargest(17).plot(kind='bar')
         plt.show()

In [ ]:  # selecting features and spliting the data in 70:30 ratio
         X=shopper_data_f.drop(['SpecialDay','VisitorType','Weekend','Revenue'],axis=
         y=shopper_data_f.Revenue
         X_train, X_test, y_train, y_test = train_test_split(X, y,train_size=0.7,rand
```

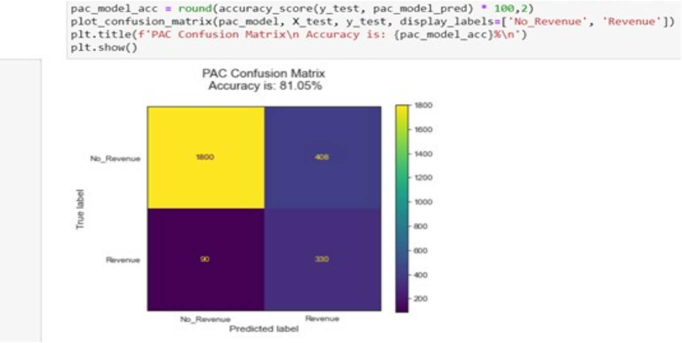
Fig. 17. Feature selection

In addition we split the split the dataset in ratio of 70:30 for training and testing sample respectively to make it ready for data modelling.

VIII. Data Visualization and Results Report

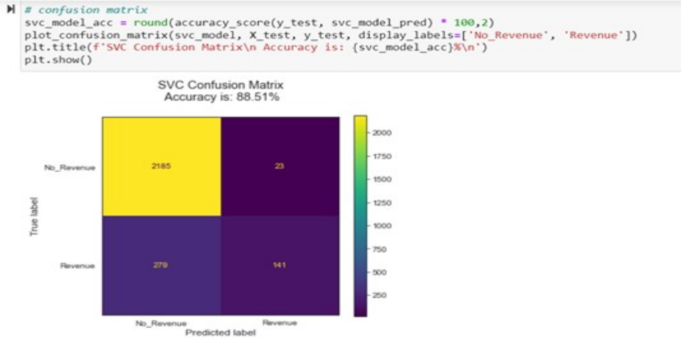
Online learning algorithms can be described as passive-aggressive algorithms. In the case that the classification is performed correctly, this kind of algorithm does not make any changes or updates; nevertheless, it becomes more active in the event that the computation is incorrect. This algorithm, in 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.

For both classification and regression, the general supervised learning approach used is SVM. In this method the data points are classified in a plane by splitting a hyper plane and the classification is done in such a way that there should be the maximum distance between the two groups. SVM will not be suitable for more than three variables as classification become harder. 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.



```
pac_model_acc = round(accuracy_score(y_test, pac_model_pred) * 100,2)
plot_confusion_matrix(pac_model, X_test, y_test, display_labels=['No_Revenue', 'Revenue'])
plt.title(f'PAC Confusion Matrix\n Accuracy is: {pac_model_acc}%\n')
plt.show()
```

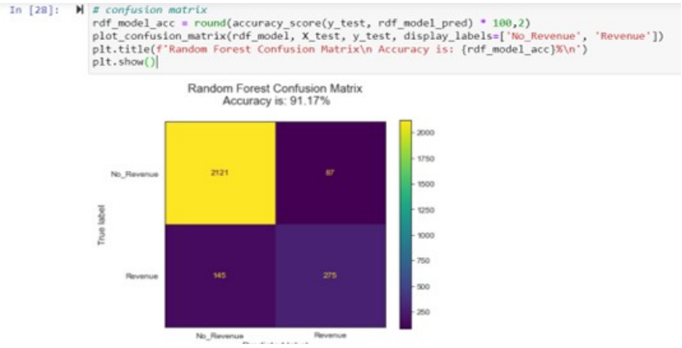
Fig. 18. Passive Aggressive Classifier



```
# confusion matrix
svc_model_acc = round(accuracy_score(y_test, svc_model_pred) * 100,2)
plot_confusion_matrix(svc_model, X_test, y_test, display_labels=['No_Revenue', 'Revenue'])
plt.title(f'SVC Confusion Matrix\n Accuracy is: {svc_model_acc}%\n')
plt.show()
```

Fig. 19. SVM

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).



```
In [28]:  # confusion matrix
          rdf_model_acc = round(accuracy_score(y_test, rdf_model_pred) * 100,2)
          plot_confusion_matrix(rdf_model, X_test, y_test, display_labels=['No_Revenue', 'Revenue'])
          plt.title(f'Random Forest Confusion Matrix\n Accuracy is: {rdf_model_acc}%\n')
          plt.show()
```

Fig. 20. Random Forest Classifier

IX. 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. Random forest model is selected since it outperforms the other models.

References

[1] 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

[2] Young Kim E Kim Y. "Predicting online purchase intentions for clothing products". In: European Journal of Marketing 38.7 (2004), pp. 883–897