## 1. Introduction

Due to the pandemic, many businesses have shifted to providing online shopping services, driving significant growth in e-commerce. To stay competitive, companies are investing in advanced technologies like machine learning. One key application is predicting Online Shopper Purchase Intent (OSPI), which helps determine whether a customer will complete a purchase. This technology is essential for understanding shopper behavior, maintaining customer engagement, and optimizing marketing strategies to improve conversions and sales. By predicting customer actions, businesses can create personalized experiences that drive satisfaction and loyalty.

## 2. Objective

The objective of this project is to analyze the “Online Shopping Purchase Intention Dataset” and explore how different features influence purchasing decisions. The study aims to develop and evaluate classification models to predict whether a website visit results in a purchase, categorized as either “TRUE” or “FALSE.”

## 3. Gathering Data

The dataset comprises 12,330 samples with 10 numerical and 7 categorical features, alongside a binary target variable, Revenue:

- FALSE: Website visit did not result in a purchase.
- TRUE: Website visit resulted in a purchase.

  Source: UCI Machine Learning Repository.

Key Attributes:

1. Administrative, Informational, Product-Related Pageviews: Pageviews grouped into specific categories.
2. PageValues: Contribution of specific pages or groups to revenue.
3. Special Day: Indicates if the visit occurred on a holiday.
4. Month, Operating System, Browser: Session-related information.
5. Visitor Type: Nature of the visitor.
6. Region, Traffic Type: Geographical and traffic details.
7. Bounce Rate, Exit Rate: Metrics indicating user engagement.
8. Weekend: Whether the visit occurred on a weekend.
9. Revenue: A binary indicator of whether the website visit resulted in a purchase. This is the target variable with binary attributes 0 & 1.

### 3.1. Summary Statistics

The summary statistics help us grasp the nature of our dataset, guiding us towards informed decisions during data preprocessing and model selection.

- Features like 'Administrative,' 'Administrative_Duration,' and 'Informational' have varying ranges and distributions.
- 'BounceRates' and 'ExitRates' also show a range of values, indicating user engagement.
- 'Weekend' and 'Revenue' are boolean (logical) variables.

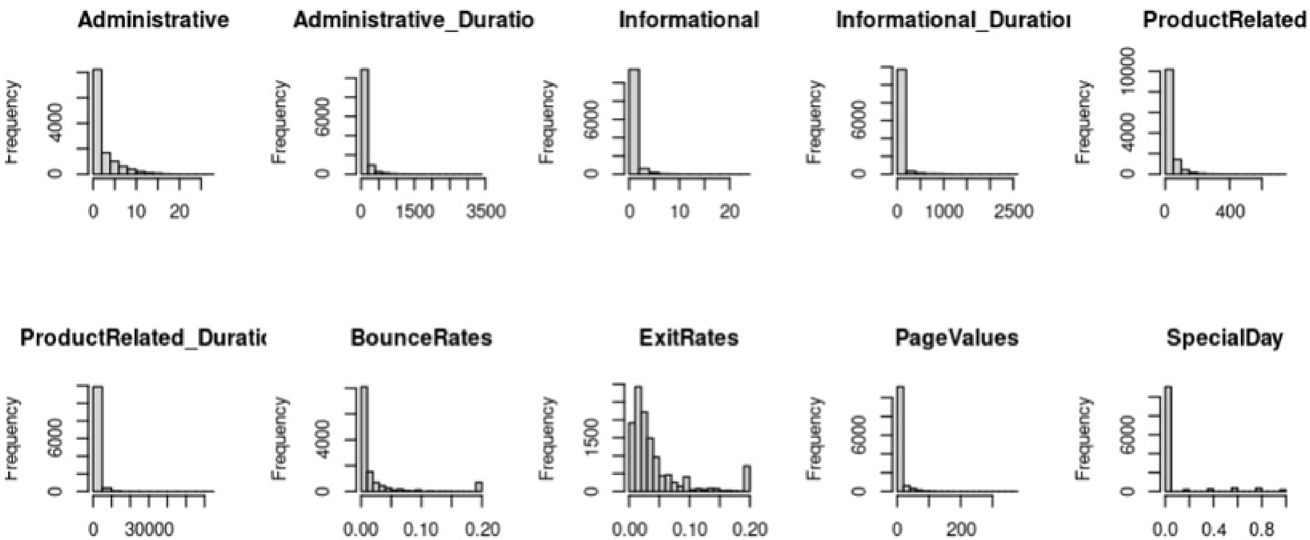
```
Administrative   Administrative_Duration Informational     Informational_Duration
Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000   Min.   :   0.00
1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000   1st Qu.:   0.00
Median : 1.000   Median :   7.50         Median : 0.0000   Median :   0.00
Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036   Mean   :  34.47
3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000   3rd Qu.:   0.00
Max.   :27.000   Max.   :3398.75         Max.   :24.0000   Max.   :2549.38

ProductRelated   ProductRelated_Duration  BounceRates        ExitRates
Min.   :  0.00   Min.   :    0.0         Min.   :0.000000   Min.   :0.00000
1st Qu.:  7.00   1st Qu.:  184.1         1st Qu.:0.000000   1st Qu.:0.01429
Median : 18.00   Median :  598.9         Median :0.003112   Median :0.02516
Mean   : 31.73   Mean   : 1194.8         Mean   :0.022191   Mean   :0.04307
3rd Qu.: 38.00   3rd Qu.: 1464.2         3rd Qu.:0.016813   3rd Qu.:0.05000
Max.   :705.00   Max.   :63973.5         Max.   :0.200000   Max.   :0.20000

  PageValues        SpecialDay           Month       OperatingSystems    Browser
Min.   :  0.000   Min.   :0.00000   May    :3364   2      :6601     2      :7961
1st Qu.:  0.000   1st Qu.:0.00000   Nov    :2998   1      :2585     1      :2462
Median :  0.000   Median :0.00000   Mar    :1907   3      :2555     4      : 736
Mean   :  5.889   Mean   :0.06143   Dec    :1727   4      : 478     5      : 467
3rd Qu.:  0.000   3rd Qu.:0.00000   Oct    : 549   8      :  79     6      : 174
Max.   :361.764   Max.   :1.00000   Sep    : 448   6      :  19     10     : 163
                                    (Other):1337   (Other):  13     (Other): 367
    Region      TrafficType               VisitorType     Weekend       Revenue
1      :4780   2      :3913   New_Visitor      : 1694   FALSE:9462   FALSE:10422
3      :2403   1      :2451   Other            :   85   TRUE :2868   TRUE : 1908
4      :1182   3      :2052   Returning_Visitor:10551
2      :1136   4      :1069
6      : 805   13     : 738
7      : 761   10     : 450
(Other):1263   (Other):1657
```

### 3.2. Data Cleaning

To enable numerical analysis, the categorical variable 'Month,' representing the month of online visits, was converted into numerical codes. Similarly, the target variable 'Revenue,' which indicates purchase intent, was transformed into a factor with levels "FALSE" and "TRUE" to make it suitable for classification tasks. Additionally, to ensure compatibility and accuracy in the Random Forest model, the ‘make.names’ function was used to maintain consistent factor levels between the predicted values and the actual data. These steps collectively ensured the dataset was clean, well-structured, and ready for effective analysis and modeling.

### 3.3. Data Preprocessing

The histograms revealed a non-normal distribution of continuous variables, indicating the need for preprocessing steps to address this issue. Feature transformations were applied as necessary to normalize the data and ensure effective model training, enhancing the accuracy and reliability of the classification models.




Here, we can see that the data used does not follow normal distribution.

### 4. Classification

Classification is a fundamental concept in machine learning that focuses on categorizing data into predefined classes or groups based on observed patterns and features. It is widely applied in various fields, including marketing, healthcare, and finance, to make data-driven decisions. Numerous classification models are available, such as Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), and Random Forests, each offering unique approaches to identifying patterns in data. For this dataset, three models were selected for evaluation: Decision Tree, KNN, and Logistic Regression Model. These models were chosen for their effectiveness in handling complex datasets and their ability to provide meaningful insights for classification tasks.

## 5. Model Selection and Evaluation

### 5.1. Decision Tree Model

Decision Trees are intuitive and powerful classification models that make predictions by repetitively dividing data into subsets based on features

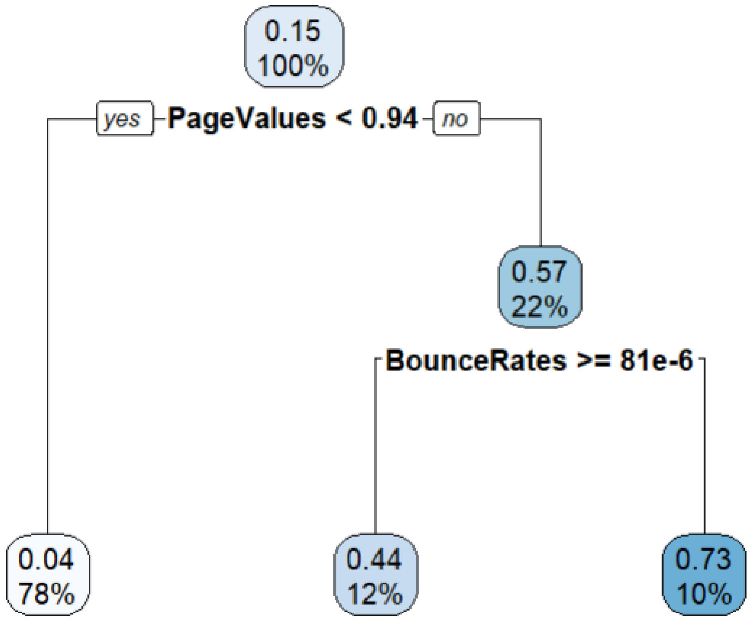
**Prediction using Decision tree model**

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10107  1035
         1   315   873

               Accuracy : 0.8905
                 95% CI : (0.8849, 0.896)
    No Information Rate : 0.8453
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5052

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9698
            Specificity : 0.4575
         Pos Pred Value : 0.9071
         Neg Pred Value : 0.7348
             Prevalence : 0.8453
         Detection Rate : 0.8197
   Detection Prevalence : 0.9036
      Balanced Accuracy : 0.7137

       'Positive' Class : 0
```

The decision tree model has an accuracy of 89.05%, with a high sensitivity of 96.98% for detecting the majority class (Revenue = 0). It performs better than logistic regression in detecting the minority class (Revenue = 1), with a specificity of 45.75%. The Kappa value of 0.5052 indicates moderate agreement, and the model shows a positive predictive value of 90.71% and a negative predictive value of 73.48%. Overall, it performs well for the majority class but could improve in identifying the minority class.

**Tree plot from a single decision tree model**

The goal of decision tree is to choose the feature that results in the best separation of the target variable (Revenue). Here the decision nodes are PageValues and BounceRates.




### 5.2. KNN Model

k-Nearest Neighbors (KNN) is a classification technique that makes predictions based on the similarity of data points in the feature space.

**Prediction using KNN Model**

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10215  1123
         1   207   785

               Accuracy : 0.8921
                 95% CI : (0.8865, 0.8976)
    No Information Rate : 0.8453
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4871

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9801
            Specificity : 0.4114
         Pos Pred Value : 0.9010
         Neg Pred Value : 0.7913
             Prevalence : 0.8453
         Detection Rate : 0.8285
   Detection Prevalence : 0.9195
      Balanced Accuracy : 0.6958

       'Positive' Class : 0
```

The KNN model has an accuracy of 89.21%, with a sensitivity of 98.01% for detecting the majority class (Revenue = 0). Its specificity is 41.14%, showing a moderate ability to identify the minority class (Revenue = 1). The Kappa value of 0.4871 indicates moderate agreement, similar to the decision tree model. The positive predictive value is 90.10%, and the negative predictive value is 79.13%. Overall, the KNN model performs well for the majority class but needs improvement in detecting the minority class.

```
k-Nearest Neighbors

12330 samples
   17 predictor
    2 classes: '0', '1'

Pre-processing: centered (18), scaled (18)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 11097, 11096, 11097, 11097, 11098, 11097, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.8748581  0.4213965
  7  0.8781016  0.4223937
  9  0.8800490  0.4258860

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 9.
```

The performance slightly improved as k increased, with the best results achieved at k = 9, showing an accuracy of 88.00% and a Kappa of 0.43. The Kappa values indicate moderate agreement between predicted and actual values.

### 5.3. Logistic Regression  Model

Logistic regression is a statistical model used for binary classification, predicting the probability of an outcome (0 or 1) based on a linear combination of input features. It outputs probabilities between 0 and 1 using the sigmoid function and is commonly applied in scenarios like predicting customer behavior or disease presence.

**Prediction using KNN Model**

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10184  1202
         1   238   706

               Accuracy : 0.8832
                 95% CI : (0.8774, 0.8888)
    No Information Rate : 0.8453
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4375

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9772
            Specificity : 0.3700
         Pos Pred Value : 0.8944
         Neg Pred Value : 0.7479
             Prevalence : 0.8453
         Detection Rate : 0.8260
   Detection Prevalence : 0.9234
      Balanced Accuracy : 0.6736

       'Positive' Class : 0

>
```

The logistic regression model has an accuracy of 88.32%, with high sensitivity (97.72%) for class 0 (Revenue = 0), but lower specificity (37.00%) for class 1. The Kappa value of 0.4375 suggests moderate agreement. While it performs well in predicting class 0 (89.44% positive predictive value), it struggles with class 1 (74.79% negative predictive value). The model's high sensitivity reflects the dominance of class 0 in the dataset (84.53%).

## 6. Conclusion

In terms of accuracy, the Decision Tree model (89.05%) outperforms both the Logistic Regression model (88.32%) and the k-Nearest Neighbors (KNN) model (89.21%). However, the KNN model has the highest accuracy, albeit slightly better than the Decision Tree. While all models show good performance, the Decision Tree and KNN models exhibit higher specificity for identifying the minority class (Revenue = 1) compared to Logistic Regression. The Logistic Regression model excels in sensitivity for classifying the majority class (Revenue = 0), but struggles with class 1. Overall, KNN and Decision Tree provide slightly better overall accuracy, while Logistic Regression has higher sensitivity but lower specificity.