Online Shoppers Intention

Problem Statement:

Based on given data of visitors browsing for online shopping, build different clusters to know whether person is only browsing and visiting multiples pages or also generating revenue for the shoppers as well. Analyse and compare with the existing Revenue Column.

Data Set Information:

The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

Attribute Information:

The dataset consists of 10 numerical and 8 categorical attributes.
The 'Revenue' attribute can be used as the class label.

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories.

The values of these 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. The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. 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. 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. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. The "Special Day" feature indicates the closeness of the site visiting time to a specific 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.

Expected Approach/Outcomes:

To study the characteristics of the features provided in the dataset and to find whether the session generates revenue or not.

Techniques used:

- **Power Transformation** was used to treat high skewness.
- Many ML Algorithms were compared but Gaussian Naïve Bayes and XG Boost gave the best results.

Outcome: Recall Score of 83% was achieved by using Gaussian Naive Bayes algorithm.

Factors affecting Revenue Generation was found (Increase in Page Value, Decrease in Bounce Rate and Exit Rate and reducing Ad duration were main factors for generating Revenue).

Key skills: Data Preparation | Feature Extraction | Sampling Techniques | Non – Linear Models Building.