Assignment Decision Trees

Total Marks : 35

Bullet point 4 to 8 – 2 points each , 10 point

Bullet point 9 – 5 point

Bullet point 10 – 5 point

Bullet point 11 – 5 point

Bullet point 12 to 17 - total 10 points

Online Shoppers Purchasing Intention

Typically e-commerce datasets are proprietary and consequently hard to find among publicly available data. This is a transactional data set that contains all the transactions occurring in online retail. Ecommerce data contains information relating to the visitors and performance of an online shop. It's mostly used by marketers e.g. in understanding consumer behavior and enhancing conversion funnels.

Objective

The objective is to build a model to predict whether a customer will buy a product or not.

Dataset

The data contains information on web sessions of a customer:

- "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration": These 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.
- Bounce Rate: 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.
- Exit Rate: The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that was the last in the session.

- Dataset has average bounce rates and exit rates for a page customer landed on.

- Page Value: The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

- Special Day: 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 the 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.

- The dataset also includes the operating system, browser, region, traffic type - these values are masked.
- VisitorType: returning visitor, new visitor, or other types of customer.
- weekend: a Boolean value indicating whether the date of the visit is weekend or not
- month: month of the year

1. import the necessary libraries

2. Read the Dataset

3. View the first and last five rows of the dataset

4. Understand the Shape of the Dataset

5. Check the duplicate data and print the number of duplicated values in each column

6. Now Drop the duplicate values

7. Check and Print the datatypes of each column . Notice that there are two columns of bool type.

We already know that we convert the categorical data-types into suitable form .

But here we have a **bool** data type. Do we have to do anything to take care of that? If yes , then what ?

8. Print the missing values in each column

9. Create a list of numerical and categorical columns

10. Print the overview of the distribution of categorical variables in a dataset, which can be helpful for understanding the data and for preparing it for machine learning models. i.e. print the

unique values in each column and proportions of each unique value in a set of categorical variables.

11. Write the insights/comment from the categorical columns. Comment on your understanding /observation from the analysis of this insights.

12. Write the insights from the numeric data.

13. Perform EDA through Data Visualization and write down any important insights gain.

14. Prepare the Data for Analysis

15. Build the model.

16. Get the Accuracy Score. Test the dataset

17. Visualize the decision tree created by the model.