Media Data Analyst- Assignment

Google Merchandise Store Problem

Understanding the problem

The attached dataset contains website users data (Google Analytics data) from Jan 1, 2017 to Jul 31, 2017.

The sample dataset contains obfuscated Google Analytics 360 data from the Google Merchandise Store, a real ecommerce store. The Google Merchandise Store sells Google branded merchandise. The data is typical of what you would see for an ecommerce website. It includes the following kinds of information:

Traffic source data: information about where website visitors originate. This includes data about organic traffic, paid search traffic, display traffic, etc.

Content data: information about the behavior of users on the site.

Transactional data: information about the transactions that occur on the Google Merchandise Store website.
Fields:

FullVisitorId: The unique visitor ID

VisitNumber: The session(visit) number for this user. If this is the first session, then this is set to 1.

Date: The date of the session in YYYYMMDD format.

VisitStartTime: The timestamp (expressed as POSIX time)

totals_bounces: Total bounces (for convenience). For a bounced session, the value is 1, otherwise it is null

totals_pageviews: Total number of pageviews within the session.

totals_timeOnSite: Total time of the session expressed in seconds.

totals_totalTransactionRevenue: Total transaction revenue, expressed as the value passed to Analytics multiplied by 10^6 (e.g., 2.40 would be given as 2400000)

totals_transactions: Total number of ecommerce transactions within the session

trafficSource_source: The source of the traffic source. Could be the name of the search engine, the referring hostname, or a value of the utm_source URL parameter

trafficSource_medium: The medium of the traffic source. Could be "organic", "cpc", "referral", or the value of the utm_medium URL parameter.

trafficSource_campaign: The campaign value. Usually set by the utm_campaign URL parameter device_deviceCategory: The type of device (Mobile, Tablet, Desktop).

device_operatingSystem: The operating system of the device (e.g., "Macintosh" or "Windows").device_mobileDeviceModel: The mobile device model.

geoNetwork_city: Users' city, derived from their IP addresses or Geographical IDs.

ChannelGrouping: The Default Channel Group associated with an end user's session for this View

We have to build a decision tree prediction model to predict if the new visitor will transact or not. When the new visitor visits the website, we get the information about source, medium, campaign, deviceCategory, operatingSystem, city, channelGrouping, pageviews, timeOnSite, bounce, etc.

Steps for solving the problem.

- **Reading the Data**
- **Data cleaning and treatment**
- **EDA**
- **Data preperation**
- **Visualisation of the decision tree**

After the whole process the accuracy came as Accuracy: 0.9996267906869922

```
In [111]: 1 print ('Accuracy: ', accuracy_score(y_test, y_test_pred))
          2 print ('\n clasification report:\n', classification_report(y_test,y_test_pred))
          3 print ('\n confussion matrix:\n',confusion_matrix(y_test, y_test_pred))

Accuracy:  0.9996267906869922

 clasification report:
               precision    recall  f1-score   support

          0.0       1.00      1.00      1.00    137458
          1.0       0.98      1.00      0.99      1827
          2.0       0.12      0.02      0.04        41
          3.0       0.00      0.00      0.00         2
          4.0       0.00      0.00      0.00         3
          7.0       0.00      0.00      0.00         0
          8.0       0.00      0.00      0.00         1

     accuracy                           1.00    139332
    macro avg       0.30      0.29      0.29    139332
 weighted avg       1.00      1.00      1.00    139332


 confussion matrix:
 [[137458      0      0      0      0      0      0]
 [     0   1821      6      0      0      0      0]
 [     0     40      1      0      0      0      0]
 [     0      1      1      0      0      0      0]
 [     0      3      0      0      0      0      0]
 [     0      0      0      0      0      0      0]
 [     0      0      0      0      0      1      0]]
```