BOARD

Machine Learning Assignment

Guidelines:

i. ***This assignment is mandatory for everyone***

ii. ***Use the below datasets to solve the below queries.***

iii. ***It is mandatory to submit the answer with the screenshot of the output you have received.***

iv. *There will only be a single attempt for each exam and **no deadline extension in case of assignments**.*

v. *Any case of unfair means or **plagiarism would lead to debarring** in final placements without any further consideration.*

Problem 1:

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 e-commerce store. The Google Merchandise Store sells Google-branded merchandise. The data is typical of what you would see for an e-commerce 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 behaviour of users on the site.

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

Dataset: https://drive.google.com/file/d/1pDIStE2jAgH44V_1KKI37ewrYdr2pgii/view?usp=sharing

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 e-commerce 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

Schema details can also be found here.

Problem Statement:

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.

Deliverables:

- R/python code.
- Visualizations (if any e.g. graphs, EDA charts, decision tree, etc.).
- documentation/notes (steps are taken to approach the problem, assumptions if any).

Problem 2:
Dataset:
https://drive.google.com/file/d/1Mh5T16w__Eb_SwLwM_jO52LNHZHcC_PB/view?usp=sharing

1. Build a Predictive Model using Data.csv
2. Use random 60% of the data set for training and random 40% of the data set for testing
3. Handle missing values using appropriate methods
4. Write a function called “feature engineering” to create bins
5. Create an excel file which has accuracy, recall, precision at different predicted probability thresholds from 0.1, 0.2, 0.3....0.9 (using code)

PS – While evaluating your results, a higher emphasis will be placed on how you handle the data and not how much accuracy you got.

Problem 3:
Dataset:
https://drive.google.com/file/d/15IgTqmlyfvsymj8Vq1sgy5FUOb9aazlt/view?usp=sharing

A supermarket in Mumbai wants to understand a wide range of customers coming. They want to provide
customized service to them. In this regard, they want to understand the different types of customers they have.

Make use of mall_customer dataset.

1. Perform standard scaling while preprocessing data.
2. Find an optimal number of clusters by elbow method.
3. Implement k means clustering
4. For the same data, draw dendrogram to find an optimal number of clusters.
5. Implement hierarchical clustering by taking 5 clusters.