SOLUTION DOCUMENT: POTENTIAL LEADS FOR A D2C STARTUP

Approaching the problem:

The problem was to find the potential lead for the startup by identifying the leads which will buy product in next 3 months given the data containing various information regarding user's past activity including whether the user bought a product in upcoming 3 months.
Since, the given dataset included whether the lead bought the product in the past or not, it was a Supervised Learning Problem.
This was a classification problem with target variable ('buy') taking 2 values: 0 (lead won't buy the product) and 1(lead will buy the product).
The training set had 17 variables where 2 of them had dates and rest of them had integer type data.
Dates are non-acceptable form of data types for most models. So, Dates were changed to suitable data type.
Different suitable classification models were tried and the one with the highest F1 score was used.
Training dataset was split into 2 sets containing 75% and 25% (validation set) of the whole data. Validation dataset was used to tune the parameters of the model selected.
The approach with the highest F1 score on validation set was finalized and used to predict the target variable 'buy' on given test dataset.

Preprocessing the dataset:

There were two problems with the dataset provided:

1. Data type of 'created_ at' and 'signup_date' columns:
   They could not be used as they contained 'object' type data and that could not be directly changed to useful integers. But, calculating how far ago leads were dropped and signed up on the website can be a useful indicator of whether they will buy the product or not. So, all dates were changed to how many days ago the lead dropped or signed up. The days could be easily converted to integer data type.
2. Missing values in 'products_ purchased' and 'signup_date' columns:
   Missing values could indicate no product purchased in 'products_purchased' and no signup in 'signup_date'. The natural step will be to replace missing values with 0 in both the columns.

After these steps, all the features were normalized as the range of values were very different for different features given which might affect our accuracy in predictions (as also found when tested).

Final Model:

The model was trained on train dataset, parameters were tested on validation set and finally fitted to test set.
After trying different models like logistic regression, random forest classifier etc., the model that gave the highest F1 score on validation set was selected. The final model was Neural Network (Multi-layer Perceptron Classifier). Different values for the different parameter like no. of hidden units, maximum iteration were tested and the ones with highest F1 score on validation set were selected.

That was the final model which was finally applied to test set given and uploaded.