Approach:

1/ Imported all the necessary libraries

2/ Imported the train dataset

3/ Performed exploratory data analysis:

- Here we can see the shape of the dataset
  where No of Columns: 19
  No of Rows: 39161
- I have calculated the Missing count, Unique count and Percent of missing values to ch eck if
  there is any missing value.
- I also checked the descriptive statistics to see the central tendency value, min, max an d sd.
- I used some plots to understand the dataset

4/ Identify and handling the missing values:

Form the dataset we can see for signup_date and products_purchased we are having 38.59 % and 53.40 % missing value respectively.

- For products_purchased: we can see the lead has not purchased anything while dropping the lead so we have imputed this column using 0.
- For signup_date: For this we have created a categorical column signup_status out of signup_date and created_at with the domain values without removing the missing values where we have three categories already_signed_up, signup_after_camp, Others.
- For created_at: We have extracted year, month and day

5/ Checked correlation value to understand the relationship between dependent and independent variables and also to avoid multicollinearity.

6/ Separated the independent and dependent features

7/ One hot encoding of 'signup_status'

8/ Splitting the dataset into training set and test set

9/ To identify best model we have performed Logistic Regression(Ridge), Random Forest, Catboost,

XGBoost Classifier, Logistic Regression, Light Gradient Boosting models

10/ I tried using hyperparameters through gridsearch cv but base model was working better. Then same preprocessing which was done for train data frame has to be done with test dataframe, then predict using the model which gave the better score better i.e Light Gradient Boosting Classifier which is a fast, distributed, high – performance gradient boosting framework based on decision tree algorithm.
11/ compare to all other model based on f1 score we could see Light Gradient Boosting models is best fitted model.
12/ After predicting the test set I saved that on a data frame replaced with sample submission ‘buy’ column.