1st step – Importing required libraries

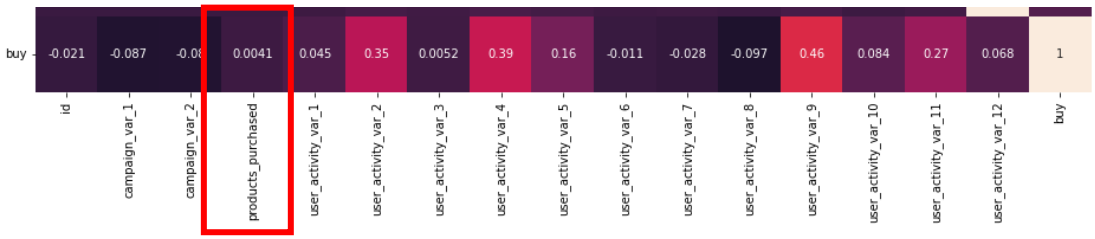
2nd step – Loading data and basic EDA

1. Info() and describe functions explains a lot about the dataset and the features.
    a. Info() shows datatypes or each variable and non-null values
    b. Describe lists all stats of all variable by looking at mean and median we can tell is there any skewness in the data or not since the difference between mean and median is less for most of columns, so we conclude that very less skewness is there and it will not affect much. So we can proceed further.
2. By isnull ans sum function its clear that 2 columns have null values present

```
df.isnull().sum() # now printing null values count
```

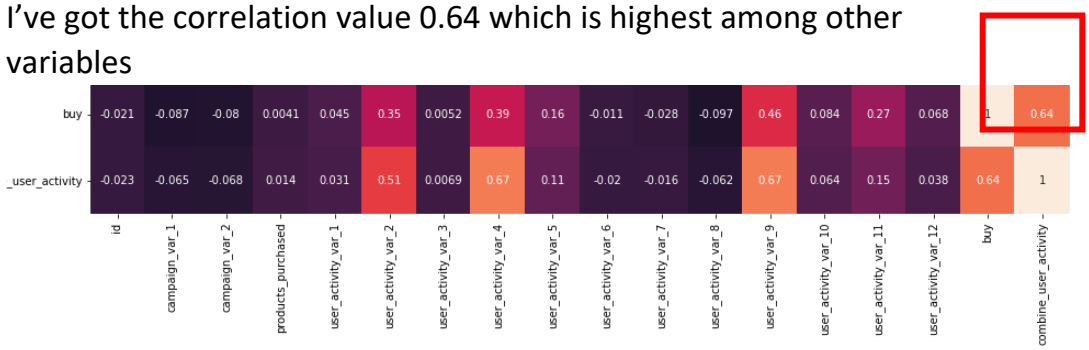
```
id                          0
created_at                  0
campaign_var_1              0
campaign_var_2              0
products_purchased      20911
signup_date             15113
user_activity_var_1         0
user_activity_var_2         0
user_activity_var_3         0
user_activity_var_4         0
user_activity_var_5         0
user_activity_var_6         0
user_activity_var_7         0
user_activity_var_8         0
user_activity_var_9         0
user_activity_var_10        0
user_activity_var_11        0
user_activity_var_12        0
buy                         0
dtype: int64
```

   But instead of directly removing those null values I plotted correlation plot to check the does there columns really have major role in predicting our target value.
3. In correlation plot we can see the columns which have null values have very less correlation with target variable and we can ignore them



3rd Step –

Feature engineering – Now after seeing the correlation plot i thought to combine few features and make single column so that it will be easy for the model as well and also we can get feature with even higher correlation. I have written python code for the same.

a. I've combined user_activity_var_2, user_activity_var_4 and user_activity_var_9 to produce single column name combine_user_activity
b. I've used simple approach to make column value as 1 if any of these above mention column is 1 else 0
c. I've got the correlation value 0.64 which is highest among other variables
d. 

4th Step –

Feature selection – Now for model building I've selected features with high correlation to avoid training the model with noise. I've selected the following features

```
[11] # feature selection - selecting only features with high correlation
     X= df[['user_activity_var_5','user_activity_var_10','user_activity_var_11','combine_user_activity']]
     Y= df[['buy']]
```

5th Step –

Data splitting – I've splitted data in ratio of 20 : 80 for testing and training respectively.

6th Step –

Model selection – So I've tried different models and at last selected decision tree classifier since its producing better results.

After model selection I've initialized the model and trained using training data

7th step –

Model evaluation – so it's giving accuracy of **97.5 %** on test data and precision of **93%**.

8th step –

Generating solution file

Loading the testing data given on the site to produce solution file.

So I've performed the same operation of feature engineering and feature selection as we did before model training

Predicted the values on testing data set.

Mapped the id and predicted value as sample submission file and saved in csv file.