Lead Scoring Approach

This Analysis is done for D2C startup and finding ways to get more cutomer to buy their products. Finding the lead who likey to buy theri products is the objective of this analysis

The following are the steps used:

1. Cleaning the Data:

- The data contained around 18 columns and around 39000 rows
- The data has 2 capaign activiy columns, date columns and user activities columns
- except the capmaign actvity and date columns all the other columns i categorical in nature
- The `products_purchased` and `sighnup_date` columns has more than 35% null values and these are the only null values columns since there are lot of information loss in these column those columns were dropped

2. Data Transformation:

- Month and Day columns were created from the `created_at` column
- The customer buying trends are increasing in nature over increase in months
- While experimenting with the various models eith various methodology, these two columns seems to be increase the Model Complexity so for model building these two columns and `created_at` column were not considered

3. Data Preparation for Model building:

- Intially the training data splittied into 70:30 ratio and various models were buit and tuned the model
- Later on the models were performing well with the entire training data being trained by the model
- So for the final solution the model being built on the entire training data

4. Model Building:

- For Basic Logistic Regression the data were scaled and statiscally checked all the variabls significance and multicollinearity checked using VIF some of the variable were dropped like `Month, Day, User_activity_9 & 10 etc.`
- Later on model building done with XGBoost, Naive Bayes, AdaBoost and Random Forest with hyperparameter tuning
- AdaBoost model performed well compared to other models

5. Hyperparameter Tuning:

- Hyperparameter tuning carried out to all the tunable models and best parameter were chosen
- Hyperparameter being tuned at finest level of granularity

6. Model Evaluation:

- All the model performance listed below,

Model	F1_Score
Logistic Regression	0.446107784431138
XG Boost without tuning	0.703488372093023
XG Boost with tuning	0.718100890207715
XGBoost with tuning and without the Month, Day and created_at columns	0.732142857142857
Naive Bays	0.61437908496732
XG Boost with tuning without month,Day, and created_at with entire training data	0.735042735042735
AdaBoost with tuning without month, day and created_at with entire training data	0.73900293255132
RandomForest with tuning without month, day and created_at with entire training data	0.725663716814159

6. Prediction on the Test set:

- With the AdaBoost model prediction has been done with the unseen test data