**Problem Statement**

To predict the propensity to buy a product based on the user's past activities and user level information.

**Data Description**

**Training set**

train.csv contains the leads information of last 1 year from Jan 2021 to Dec 2021. And also the target variable indicating if the user will buy the product in next 3 months or not.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique identifier of a lead |
| created\_at | Date of lead dropped |
| signup\_date | Sign up date of the user on the website |
| campaign\_var (1 and 2) | campaign information of the lead |
| products\_purchased | No. of past products purchased at the time of dropping the lead |
| user\_activity\_var (1 to 12) | Derived activities of the user on the website |
| buy | 0 or 1 indicating if the user will buy the product in next 3 months or not |

**Test set**

test.csv contains the leads information of the current year from Jan 2022 to March 2022. You need to predict if the lead will buy the product in next 3 months or not.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique identifier of a lead |
| created\_at | Date of lead dropped |
| signup\_date | Sign up date of the user on the website |
| campaign\_var (1 and 2) | Campaign information of the lead |
| products\_purchased | No. of past products purchased at the time of dropping the lead |
| user\_activity\_var (1 to 12) | Derived activities of the user on the website |

**Data Preprocessing**

1. Checking and analysing the description of the data.
2. As we are dealing with train and test set we need to check both datasets for any missing values.
3. We have too many missing values in the products purchased column, but it seems this column is valuable that is why we will fill the null values with zeroes.
4. The data has signup\_date (date at which customer signed up on the website) and created\_at (date at which the lead was dropped), we can create a new attribute which will contain the total days during which the lead was held.
5. Since the date columns is in float type we need to change it into the date time columns to apply our function to create our new attribute (created\_at – signup\_date).
6. Change back the data type to float.
7. Creating a lambda function to remove the ‘days’ part from the new attribute would help us building our model more efficient.
8. Dropping the dates and id columns as we enter into model training part.

**Model Training**

1. Splitting the dataset into training and validation set to validate our model.
2. We will use different models to see how they perform without any tuning.
3. We observe that Random Forest performs decently on the model now we will tune the hyper parameters by finding the best parameters using RandomSearchCV.
4. After getting the best parameters we will re-train our Random Forest on the selected parameters.
5. We see that the model score is slightly less, this would be because we used Random Search instead of Grid Search due to my system limitations.
6. We will now check how xgboost works on our data.
7. As we can see that xgboost performs better than any of the models, let’s see if we can increase the performance by tuning the hyper parameters.
8. We can absolutely increase the performance of our xgboost using the hyper parameter tuning.
9. Finally, we submit our prediction on test data.