Approach

**Brief:**

We have two datasets names train & test. Train dataset contains  the leads information of last 1 year from **Jan 2021 to Dec 2021**. And the **target variable** indicating if the user will buy the product in next 3 months or not. Test. dataset 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**.

As test dataset don’t have labels, we need to predict labels from he model trained on train dataset. As this is classification machine learning. We use RandomForestClassifier algorithm Using Scikitlearn. Ensemble module. We can split train dataset into training data and validating data. We have to standardize data for modeling data. We train model on training data and validate on validating model. Now we can predict model on test data.

**Data Preprocessing:**

First, we read data on pandas dataframe using pandas. We did basic exploration of data by exploring number of rows & columns by pandas. We have 39161 rows & 19 columns in train dataset. We have 13184 rows with 18 columns in test dataset. We explore names of columns by pandas. We have 'id', 'created at', 'campaign\_var\_1', 'campaign\_var\_2','products\_purchased', 'signup date', 'user\_activity\_var\_1','user\_activity\_var\_2', 'user\_activity\_var\_3', 'user\_activity\_var\_4','user\_activity\_var\_5', 'user\_activity\_var\_6', 'user\_activity\_var\_7', 'user\_activity\_var\_8', 'user\_activity\_var\_9', 'user\_activity\_var\_10','user\_activity\_var\_11', 'user\_activity\_var\_12', 'buy' in train dataset.

We have same columns except buy in test dataset. We need to find buy values in test by prediction. We must explore data is duplicated or not. We need to check missing values. We have missing values in products purchased, signup date. Signup date is categorical data, we need to replace null values with mode. We need to check distribution of products purchased variable. It shows symmetrical, so null values replaced by its mean. We need to convert signup date, created at to timestamp datatype using astype function. We need to convert products purchased to int datatype using astype. We have id attribute, which is unique number, so we can convert this into category. So, we don’t need to calculate mean & median to it. We have o calculate correlation for all numerical variables. It didn’t show any correlation. So, we can’t select one variable as best for modelling. We need to take all attributes to modelling.

We must predict buy values. So, in train dataset, here dependent variable is buy(Y), independent variables are other variables in data except datetime & id. In test dataset X is given data except timestamp variables. We have done EDA by plotting line plot for products purchased to signup date, created at. It shows trend of how many products purchased during that time. From 2019 to 2010 it shows decrease in number of products ranges purchased.

We must import preprocessing from sklearn library to do preprocessing & modelling data. From sklearn. preprocessing we need to import standard scalar for standardizing data. Using train\_test\_split we have to split train data into training & validating data (20%). Here sklearn worked for preprocessing data.

**Final Model:**

We have train & validate data in training data & test data. So Random Forest classifier is best to use. We must apply RFC algorithm to train data. We must validate model with validate data. We will predict model using test data. We predicted values of buy using predicting model on test data. We will find accuracy for RFC. It shows 92% accuracy, which is good model. We will calculate f1 score using sklearn metrics for different averages. For micro average f1 score is 0.92 for class label 1. Which is good.

**Conclusion:**

Here we have 2 datasets of train & test separately, in which train has label of buy, test doesn’t have, we need to predict. we read data into pandas dataframe. we did basic exploration of data by exploring shape, columns, missing values. Data munging done by converting data types of created at, signup date to timestamp. fill null values with respective mean & median respective of variable symmetric & skewed data. we examine trend of products purchased for signup date & created at. we can say it is decreasing from 2019 to 2021. so we need to focus on potential customers or lead, to increase short term sales for next 3 months. we need to predict who will likely to purchase products in next 3 months from their past data. we need to standardize data for machine learning. As we don’t have label for test, we used RF classifier & calculate accuracy of train data. here train data splits into training & validating test. This validating model used to predict label for test. Here we got accuracy of RF classifier is 92.35, which is good model. F1\_score with micro average for buy label 1 is 0.92. we created buy label column in test dataframe. we created new dataframe name buy\_1 for label 1 values along with id saved into csv format.