Introduction

A D2C startup develops products using cutting edge technologies like Web 3.0. Over the past few months, the company has started multiple marketing campaigns offline and digital both. As a result, the users have started showing interest in the product on the website. These users with intent to buy product(s) are generally known as leads (Potential Customers).

Leads are captured in 2 ways - Directly and Indirectly.
Direct leads are captured via forms embedded in the website while indirect leads are captured based on certain activity of a user on the platform such as time spent on the website, number of user sessions, etc.

Now, the marketing & sales team wants to identify the leads who are more likely to buy the product so that the sales team can manage their bandwidth efficiently by targeting these potential leads and increase the sales in a shorter span of time.

Now, as a data scientist, your task at hand is to predict the propensity to buy a product based on the user's past activities and user level information.

The list of feature is given below:

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

Initial ideas:

- **Classification model:** since the buy feature is not continuous variable buy feature classify the discrete value, one could use a classification model.
- Data is highly imbalanced**.** So, I'll perform imbalanced classification.
- **Imbalanced Classification** refers to classification where number of examples in each class is unequally distributed.
- There arore some specialized technique may be used to change the composition of samples in the training dataset by undersampling the majority calss or over sampling the minority class.
  Examles include:
  Random Undersampling.
  SMOTE Oversampling.
- Specialized modelling algorithm may be used that pay more attention to the minority class when fitting The model on the training dataset ,such as cost-sensitive machine learning algorithm.
- Finally, alternative performance metrics may be required as reporting the classification accuracy may be misleading.
  Examples include:
- **Precision**.
- **Recall**.
- **F-Measure.**

Final Model:

Checking data is clean or not:

- 2 columns have missing values.
- First column is Product_purchased, which has 53% missing value and it have four discrete value so I decided fillup with mean.
- Second column is sign_up date which can't affect target value so I decided to drop it.

Data is imbalanced so I decided to SMOTE upsampling of minority class for balncing the data without loosing any information.

After balancing the data, fit the model first I'll fit Logistic regression, but it'll not perform better on data and not give great accuracy then I'll will fit XGboost model then it will perform better. So I'll go with xgboost.

- F1 score of class 1 from xgboost is: 0.98.
- Accuracy of test data is 97%

Ideas for Improvement:

- Use of neural network for balancing the data