**D2C Startup .. Approach**

1. **Business Understanding:**

**Business Objective:** to predict the propensity to buy a product based on the user's past activities and user level information.

**Business Constraints**: 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.

**Data Understanding:**

Id – ordinal data,

created\_at – date (object),

signup\_date – object,

campaign\_var(1 and 2) – discrete data,

products purchased - continuous data,

user\_activity\_var (1 to 12) – discrete data,

buy - o or 1 (discrete binary data.) target feature.

1. **Exploratory Data Analysis:**

* Basic descriptive statistics.
* First moment business decision, Measures of central tendency (mean, median & mode)
* Second moment business decision (variance, standard deviation)
* Third moment business decision (skewness)
* Fourth moment business decision (Kurtosis).

**Graphical Representation - Fifth moment Business Decision**

To get useful & meaningful insights from the data plots are very helpful.

**Univariate analysis**: Histogram, Count plot, Boxplot

**Bivariate analysis** : Bar plot

Correlation matrix using heatmap.

1. **Data Pre-Processing:**

**Missing Values**:

Handled missing values (NA or nan) using fillna() function as farward fill.

**Outliers Treatment:**

Handled outliers using winsorizer technic & IQR capping method used.

**Imbalanced Dataset:**

Found that data is imbalanced. Class-0 is majority class and class-1 is minority class. Class-0 is dominated so it’s an imbalanced data.

**Handled date columns**:

In the given dataset 2 date columns are found and those are split by ‘-‘ and separated Date Columns as year, month and day as 3 columns & after that dropped the date columns.

**Splitting Data :**

Split data as dependent (target) & independent (predictors)

Data is split as Train, test and 80% is for training model.

''' For imbalanced datasets Tree Based Algorithms like

Decision trees often perform well on imbalanced datasets because their hierarchical structure allows them to learn signals from both classes. In modern applied machine learning, tree ensembles like Random Forests, Gradient Boosted Trees, etc. almost always outperform singular decision trees.

Here Resampling and SMOTE is covered. '''

1. **Model Building:**

In model building part I built 3 models Decision Tree Algorithm, Random Forest and XGBoost..

For **Decision Tree** I got **accuracy 96%** and **F1 Score of Class 1 is 0.52**.

For **Random Forest** the **accuracy is 96%** and **F1 Score of Class 1 is 0.52**,

For **XGBoost** the **accuracy is 97%** and **F1 Score of Class 1 is 0.63**

For **XGBoost** got accuracy as 98%

Performance metrics like confusion matrix and Classification Report is made.

As you mentioned **evaluation metric** for this hackathon is **F1 Score of Class 1**.

So, **F1 Score of Class 1** for **Random Forest is 0.70**