# **Predicting User Purchase Behaviour on an E-commerce Platform**

**Objectives:**

The primary objectives of this hackathon are:

* Predicting whether users will make purchases within the next three months using baseline models and evaluating model performance.
* Preprocessing the data by cleansing and transforming it into a format ready for machine learning.
* Analyzing weights to address data imbalance and splitting the data into training and test sets.
* Selecting the appropriate hyper-tuning parameters for the classification model.
* Implementing baseline and classification models on oversampled data and evaluating their performance.
* Evaluating and interpreting models to identify the most effective model for this dataset.

**Understanding the Data for Logistic Regression:**

Before applying the models, it is essential to identify the features that significantly impact predicting the target variable (purchase). Visualizing features related to the purchase feature is crucial for this analysis.

**Exploratory Data Analysis before creating a Logistic Regression Model:**

Comparing the target variable with other variables helps assess their impact. Heatmaps are utilized to detect correlations between the target variable and other variables. The analysis reveals that only Campaign 1 and 2 are highly correlated.

**Data Engineering & Pre-processing:**

Preprocessing involves handling missing values and transforming features to prepare the dataset for machine learning. The purchase column's NA values are replaced with zero values, and irrelevant variables such as date and ID are dropped.

**Addressing Class Imbalance:**

Class imbalance occurs when one category significantly outnumbers the other, leading to skewed distributions. Techniques such as adjusting class weights are explored to address this issue and ensure balanced training.

**Building and Training Logistic Regression Model in Python:**

Logistic Regression, a popular classification algorithm, is implemented using the Scikit-learn library. A base model with default parameters is initially trained on the dataset. Hyperparameter tuning techniques such as Grid Search are then applied to optimize model performance.

**Model Evaluation Metrics:**

Model performance is evaluated using metrics such as F1 score, precision, recall, ROC-AUC score, and confusion matrix. These metrics provide insights into the model's ability to classify instances accurately.

**Model Evaluation:**

The final tuned model is evaluated on a separate test dataset to assess its generalization performance. Accuracy values, precision, recall, and F1 scores are calculated to quantify the model's performance accurately.

**Conclusion:**

By leveraging data preprocessing techniques, feature engineering, and hyperparameter optimization, the Logistic Regression model's performance is significantly improved. The tuned model achieves a higher F1 Score of 0.71 and ROC-AUC score of 0.88 compared to the base model's F1 Score of 0.57. These findings highlight the effectiveness of the model in predicting user purchase behavior, providing valuable insights for optimizing marketing strategies and improving customer engagement.