***Template***

* *Introduction and executive summary*
* *Problem statement*
* *Methodology*
* *Result and discusion*

**EXCUTIVE SUMMARY**

This paper presents an iDAF data science capstone project for Terragon Group. The project entails building a model to forecast the likelihood of a customer clicking an ad. Terragon’s historical marketing data was used for the data analyses and model building.

This project followed a complete data science pipeline, from exploratory data analysis (EDA) to the insights reporting, and featured operations like correlation, univariate and bivariate analysis, and dimensionality reduction before fitting the model into a Random Forest algorithm. The optimal model had a precision score of 76.3% and recall of 17.3%. A Flask API endpoint was built and tested to access the model before it was deployed on Docker to make predictions in real-time.

**INTRODUCTION**

Click-through rate (CTR) is the ratio of users who click on a specific link to the number of total users who view a page, email or advertisement. It is commonly used to measure the success of an online adverting campaign for a particular website as well as the effectiveness of email campaigns.

At Terragon, behavioral, transactional, and demographic data are analyzed to calculate the propensity of profiles to respond to an ad served. Their Machine learning algorithms can forecast the likelihood that customers will purchase, install, unsubscribe, sign up or confirm a loan by finding patterns in their past data. By more intelligently analyzing the intricacies and customer journeys of previous customers, one can better understand how leads and customers were acquired in the past, and who didn't convert in the end. Finally, through personalized targeting, Terragon can determine which prospects are similar to current customers and thus likely to convert in the end.

**PROBLEM STATEMENT**

Given the demographic and behavioral patterns of profiles during a period, the problem is to calculate the propensity of profiles to click on an ad served. Ads will be displayed to potential customers with high propensity to click on them. A model of this decision problem would allow us to know which customers to target while serving an ad and increase the Click Through Rate (CTR) of campaigns.

**METHODOLOGY**

The methods adopted in this project followed the Cross Industry Standard Process for Data Mining (CRISP-DM) according to Shearer, 2000. This comprises business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

The data used was of marketing, featuring 66,739 customers and 31 variables such as customer location, age, amount spent on data, SMS costs, etc.

As a first step, the data understanding here involved carrying out some preliminary exploration using quick operations like summary and descriptive statistics on the numerical data. The ‘label’ or target variable was then identified (in this case Click), and other variables were studied to understand their definition of the dataset.

The next step in the process was the exploratory data analysis, where the iDAF team performed univariate analysis on the individual variables, and multivariate analysis on a combination of variables to determine their informativeness in predicting the target. This was achieved using visualizations to analyze distributions, correlations, balances and imbalances, etc. in the data.

In the data preparation step, a couple of pre-processing operations were carried out to ensure that the variables of interest for the model building were suitable for use. This involved cleaning operations like handling missing values and removing outliers, and transformation operations like rescaling and normalization of numerical variables, and encoding the categorical variables (to be able to use them in model building).

Furthermore, the project pipeline led to the model building stage, adopting a Train-Test split ratio of 75:25. Several algorithms were tried to train part of the data and make predictions on the test data. Random Forest was selected as the best performing algorithm using Precision as the primary metric, and the following scores were evaluated:   
Accuracy: 0.798   
Precision: 0.763  
Recall: 0.173.   
Other algorithms that were tried are Logistic Regression, Catboost, Support Vector Machines (SVM), and Decision Tree.

Finally, a machine learning application for the model was built using FlaskAPI and deployed on Docker to enable consumption in real time. The application was able to predict whether a user would click an ad (1) or not (0) on inputting the required variables data into the engine.