**Capstone Project Submission: Email Campaign Effectiveness Prediction**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Contributor Name:** Ranajay Biswas  **Email ID:** - ranojoybiswas21@gmail.com  **Contributor Role:**   * Data collection * Exploratory Data Analysis * Data Pre-processing & Feature Engineering * Graphical Representation & Data Visualization * Modelling * Feature Importance Evaluation * Hyper parameter Tuning * Conclusion |
| **Please paste the GitHub Repo link.** |
| Github Link:- https://github.com/RanojoyBiswas/Email-Campaign-Effectiveness-Prediction  Google Drive Link:-  https://drive.google.com/drive/folders/1vTqFxHW4h-S2Z5elhUt-yeZqh9ZqzgRn?usp=share\_link |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Email campaign is a sequence of marketing efforts that contacts multiple recipients at once. Email campaigns are designed to reach out to subscribers at the best time and provide valuable content and relevant offers. Using email campaigns allows businesses to build deep and trusting relationships with their customers. Marketing through Email can make communication with clients easier and more effective. Email campaigns are a very powerful medium between a business company and its audience. It helps not only to increase sales but build brand image. Most of the small to medium business owners are making effective use of Gmail-based Email marketing Strategies for offline targeting of converting their prospective customers into leads so that they stay with them in business.  The main objective is to create a machine learning model to characterize the mail and track the mail that is ignored; read; acknowledged by the reader.  Performing Exploratory data analysis helped us understand the features and relationships that they have and their impact on the target or the client's response. We tried to find out important features.  Data is labeled and the target column being categorical, we implemented classification based machine learning algorithms to complete the prediction task.  The Email campaign data contains various types of information regarding the emails that were sent, it contains info about their customers and their responses. Checking the shape of the data, we found that it has 68353 observations and 12 columns.  Null values and outliers were treated accordingly. We checked data distributions for various features. New features were created from existing and correlated features to solve the problem of multicollinearity. We used Synthetic Minority Oversampling (SMOTE) and also random Undersampling to handle the imbalance in the target column. We compared the results after training different classifiers and found that oversampling outperforms undersampling at most occasions.  After that we checked the feature importance for the models that performed the best. The least important features were discarded and using cross validation, we performed hyper-parameter tuning. Model performance improved after tuning. The best classification algorithm we found for this problem statement was XGboost classifier with oversampled data and only the important features selected.  Since the problem statement clearly mentioned that we need to characterize the mails based on the user response, we decided to use F1 Score and AUC-ROC Score. |