Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1)Rajat Chaudhary(Rajat.25.chaudhary@gmail.com) Null value treatment, Logistic regression model, Decision tree model, Models comparison, evaluation metrics 2) Anukriti Shakyawar(shakyawaranukriti@gmail.com) Cleaning data, Multicollinearity treatment, VIF, Random Forest, Decision Tree hyperparameter tuning 3)Raman Kumar(ramank445522@gmail.com) modelling, feature scaling, visualizations, model evaluation function, XG Boost 4)Deepmala Srivastava(srvdeepmala@gmail.com) Outlier detection and treatment, One-hot encoding, feature importance, data-preprocessing, Conclusion 5)Kritisha Panda(kritishapanda57@gmail.com) EDA, RF hyperparameter tuning, XGB feature importance, Random under sampling, SMOTE
Please paste the GitHub Repo link.
Github Link:- https://github.com/ramank123/Email-Campaign-Effectiveness-Prediction
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

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.

We begin with null-value treatment and fill the null values in the features with the mean or mode of the respective feature depending on whether they are normally distributed or skewed.

Followed by EDA, we check the distributions of categorical and numerical variables w.r.t the target variable to draw actionable insights and understand the trend in the data for better processing.

Features like 'Email ID' and 'Customer Location' are dropped as they won't contribute much on deciding the target variable.

On checking the correlation, we see that 'Total images' and 'Total links are positively correlated and consequently have a high VIF.

We combine the features into 'Total_img_links' and drop the original columns. The VIFs are now balanced.

Moving on to outlier treatment, since our dataset is highly imbalanced, we remove just 5% of the outliers. Too much removal will lead to loss of information.

Followed by feature scaling and one-hot encoding for categorical variables, we begin with modelling.

To treat class imbalance, we perform random under sampling as well as SMOTE separately to check the performance of models on both the datasets. We used the following models: Logistic Regression, Random Forest, Decision Trees, Random Forest with Hyperparameter Tuning, Random Forest hyper parameter tuned with feature importance and XG Boost.

Each model is trained and tested on Random under sampled data and SMOTE data. The metrics used for comparing the models are: Accuracy, Recall, Precision, F1 score and AUC for both train and test data to check cases of overfitting.

A single function 'model_evaluation' is designed which takes the names of the models, train and test sets for Random under sampled and SMOTE data and predicts and evaluated the models for classification, visualizes results and creates a data frame that compares the models.

This is done to enhance the modularity of the code. We observed that-

SMOTE worked considerably better than Random Undersampling, it may have led to loss of information.

XGBoost Algorithm worked in the best way possible with such imbalanced data with outliers, followed by Random Forest Hyperparameter Tuned model after feature selection with F1 Score of 0.71 on the test set

.