Capstone Project Submission

Wasim Ahmad

wasim.faim@gmail.com

Contribution

- Frame work of project.
- Exploratory Data Analysis
- Data visualization
- Data Preprocessing
- Feature Engineering
- Data Imbalance Handling
- Story Telling
- Model building
- Model Evaluation
- Sample PPT
- PPT presentation.
- Presentation Video
- Technical documentation.
- Project summary template

Github link

https://github.com/wasimfaim/Email-Campaign-Effectiveness-Prediction

Short Summary:

Email marketing is a form of marketing that can make the customers on your email list aware of new products, discounts, and other services. It can also be a softer sell to educate your audience on the value of business brand or keep them engaged between purchases. In 1978, a marketing manager at Digital Equipment Corp named Gary Thuerk used this new method of direct communication to send out the first commercial email to let people know about a new product.

Problem Statement: 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.

Approach Done:

The dataset given is a dataset from Email marketing. It is Email campaign data and we have to analysis the mail that is ignored; read; acknowledged by the reader

- At first we did some basic operation to understand the dataset
- The dataset has 68353 rows and 12 columns.
- There was missing values in some features which handled by fill mean median mode.
- There is no duplicate value too thankfully.
- It had outliers also which was handled.
- We did EDA (Exploratory data analysis) on the given data.
- Then we visualized the data to see the insights and patterns of the data.
- We merged two columns as images and links are correlated and most of the values in Total_Images was 0**.**
- We did feature selection according to their importance.
- Our dependent variable was highly imbalanced which was balanced by technique random under sampling
- After preprocessing the data then we split the data into 80-20 for the train-test.
- Finally we applied ML algorithms likes Decision Tree, Random Forest, Gradient Boosting, SVM, Naive Bayes and K Nearest Neighbor etc.
- XG Boost Classifier worked the best giving a train score of 89% and test score of 81% for F1 score.

Conclusion:

- In EDA, we observed that Email_Campaign_Type was the most important feature. If your Email_Campaign_Type was 1, there is a 90% likelihood of your Email to be read/acknowledged.
- It was observed that both Time_Email_Sent and Customer_Location were insignificant in determining the Email_status. The ratio of the Email_Status was same irrespective of the demographic or the time frame the emails were sent on.

- For getting better acknowledged by the reader email should be written Email Type 2, Email source type 1 and Email campaign type 1(Email campaign type 1 are 10% ignored, 66% read and 24% acknowledged by reader)
- As the word_count increases beyond the 600 mark we see that there is a high possibility of that email being ignored. The ideal mark is 400-600. No one is interested in reading long mails !
- For modelling, it was observed that for imbalance handling Oversampling i.e. SMOTE worked way better than undersampling as the latter resulted in a lot of loss of information.
- Based on the metrics, Random Forest and XG Boost Classifier worked the best giving a train score of 99% & 90 and test score of 85% & 82% for F1 score.