Capstone Project 03

Email Campaign Effectiveness Prediction

By: Prashant Gaikwad

(Individual Project)

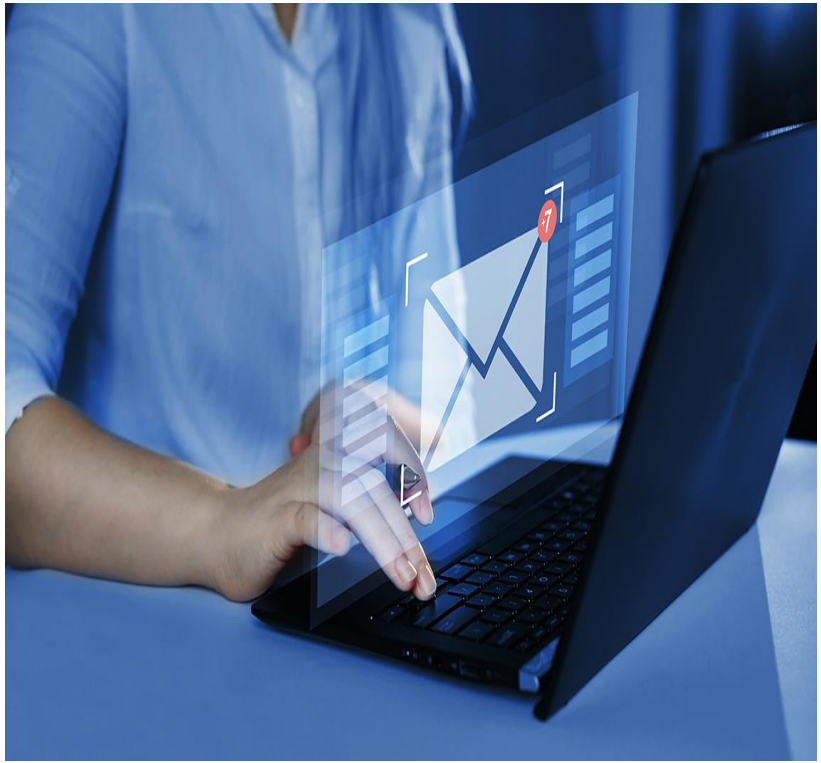
Points for Discussion

- ❑ Introduction
- ❑ Problem Statement
- ❑ Email Campaign Dataset
- ❑ Data Exploration
- ❑ Data Preparation & Cleaning
- ❑ Exploratory Data Analysis (EDA)
- ❑ Correlation
- ❑ Machine Learning Model Building
- ❑ Feature Importance
- ❑ Conclusion

INTRODUCTION

- Most of the small to medium business owners are making effective use of Gmail-based Email marketing Strategies

- We all subscribe to many different kinds of businesses through emails because it's required to do so, sometimes to get digital receipts of the things we bought or to get digital information about the business to stay updated.



- In this Project, we will see how machine learning can be used to predict **Email Campaign Effectiveness** on a real-world business problems.

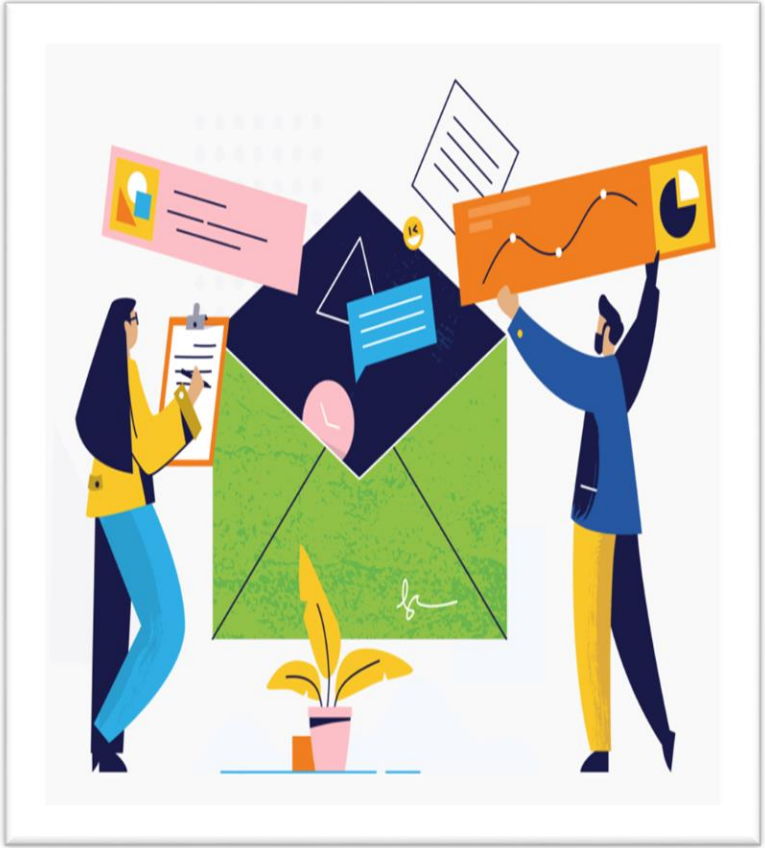
- Also, we will get to see each and every phase of how in the real world a case study is solved.

PROBLEM STATEMENT

The Problem statement is to find out the Email Campaign Effectiveness Prediction using Machine Learning algorithms.

Email marketing allows businesses to build relationships with leads, new customers and past customers. It's a way to communicate directly to the customers in their inbox, at a time that is convenient for them. With the right messaging tone and strategies, emails are one of the most important marketing channels.

We all subscribe to many different kinds of businesses through emails because it's required to do so, But many of times we do not tend to read an email due to a number of reasons - to name a few would be- no proper structure, too many images, too many links inside the mail, complex vocabulary used or simply too long emails.



Email Campaign Effectiveness Dataset

1) Email_ID

2) Email_type - Email type contains 2 categories 1 and 2. We can assume that the types are like promotional email or important email.

3) Subject_Hotness_Score - It is the email effectiveness score

4) Email_Source

5) Email_Campaign_Type

6) Total_Past_Communications - This column contains the previous mails from the same source.

7) Customer_Location - Categorical data which explains the different demographics of the customers.

8) Time_Email_sent_Category - It has 3 categories 1, 2 and 3 which may give us morning, evening and night time slots.

9) Word_Count

10) Total_Links

11) Total_Images - The banner images from the promotional email.

12) Email_Status - It is the target variable which contains the characterization of the mail that is ignored; read; acknowledged by the reader.

DATA EXPLORATION

- Data exploration refers to a data user being able to find his or her way through large amounts of data in order to gather necessary information.

- Explored and analyzed the data to discover key factors responsible for app engagement and success.

DATA PREPARATION & CLEANING

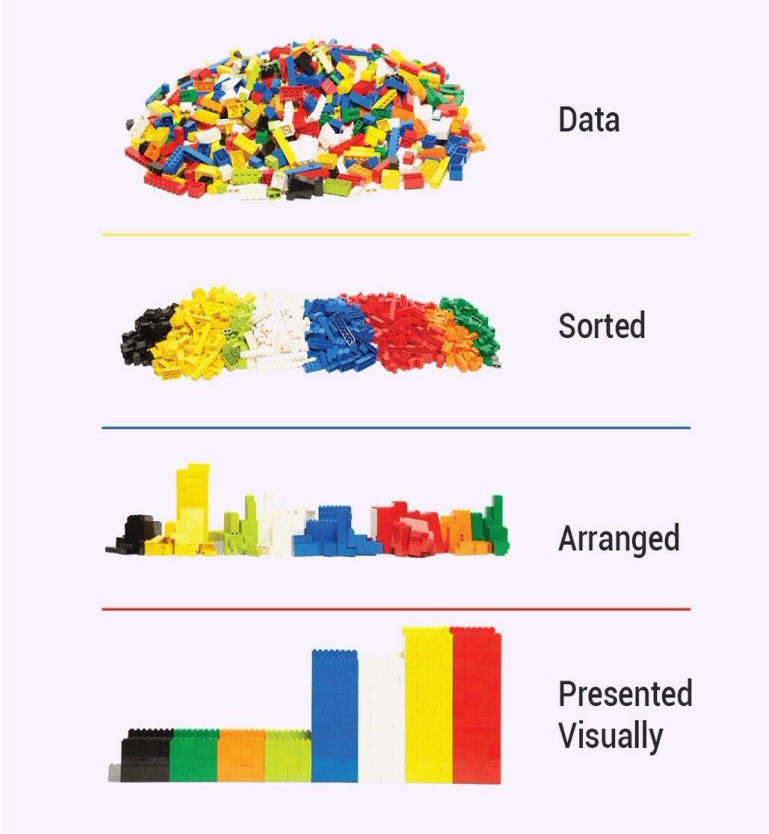
- Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and the combining of data sets to enrich data.

- Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

- Dataset contains a large number of null values which might tend to disturb our analysis hence we dropped some unnecessary variables.
- Used “fillna()” method to filled some required variable null values from them at the beginning of project in order to get a better result.
- Encoded all categorical variable to numeric values

EXPLORATORY DATA ANALYSIS

- Exploratory Data Analysis is a process of examining or understanding the data and extracting insights or main characteristics of the data.
- **EDA** is very essential because it is a good practice to first understand the problem statement and the various relationships between the data features before getting hands dirty.



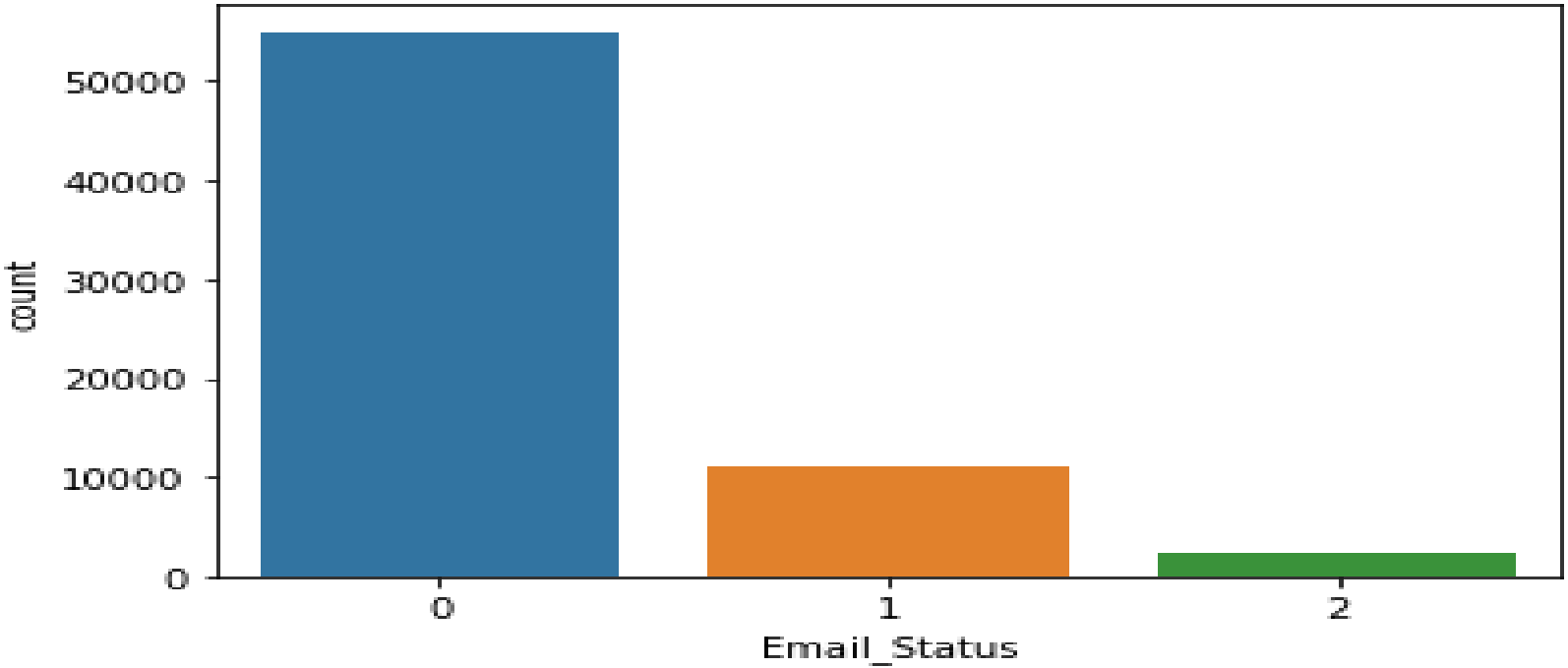
Types of EDA are:

- Univariate Analysis - analysis of a single variable
- Bivariate Analysis - analysis of exactly two variables
- Multivariate Analysis - analysis of dependent variable and multiple independent variables

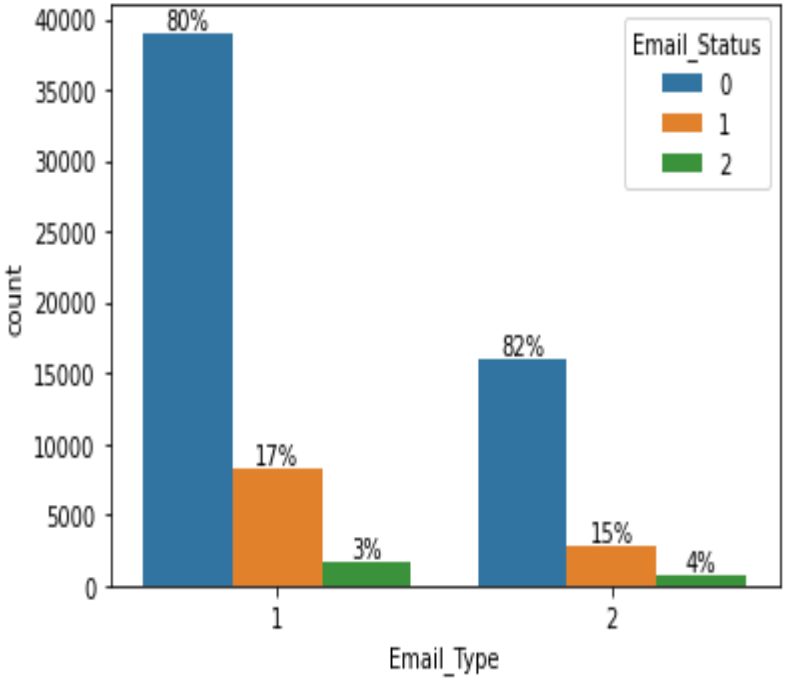
The primary motive of EDA is to

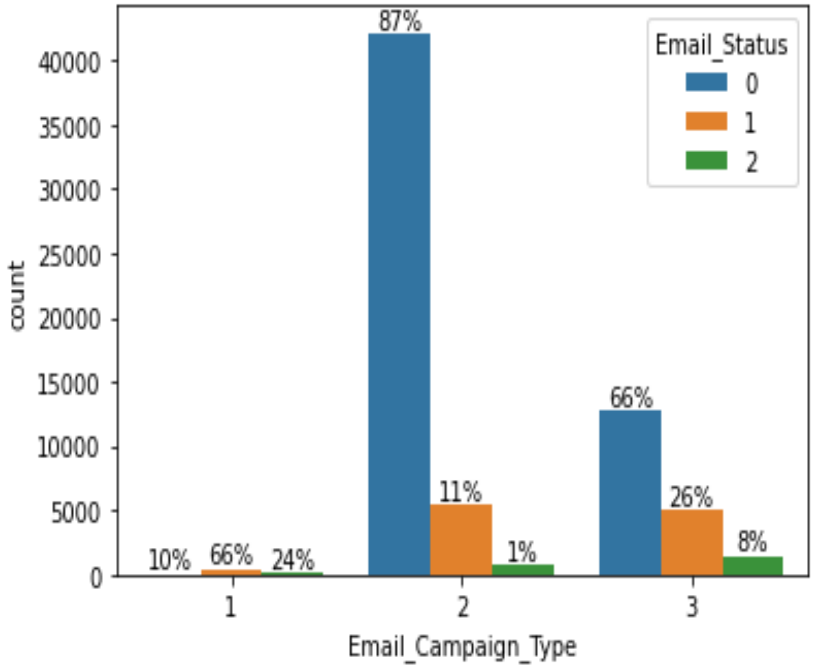
- Examine the data distribution
- Handling missing values of the dataset (a most common issue with every dataset)
- Handling the outliers
- Removing duplicate data
- Encoding the categorical variables
- Normalizing and Scaling

Email Status

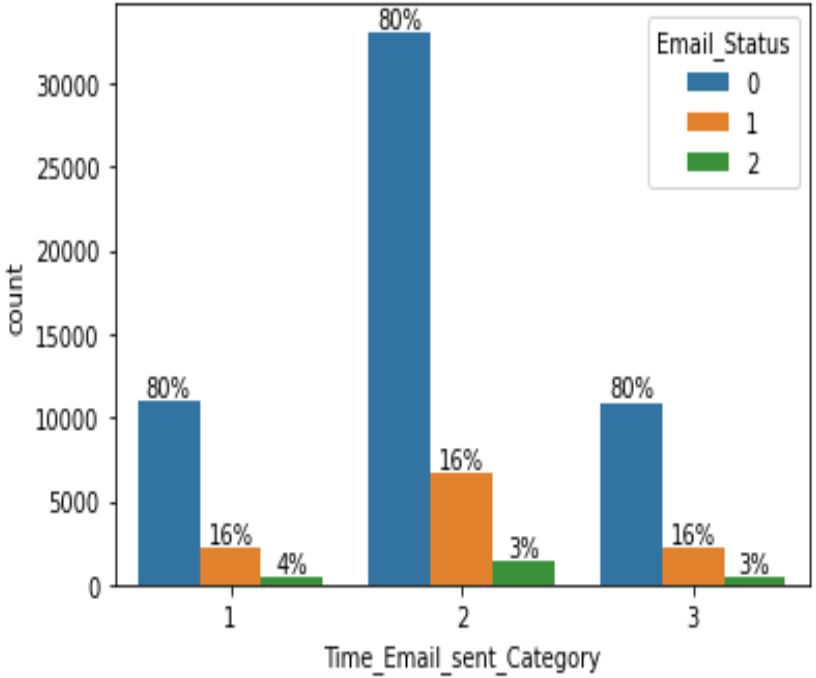


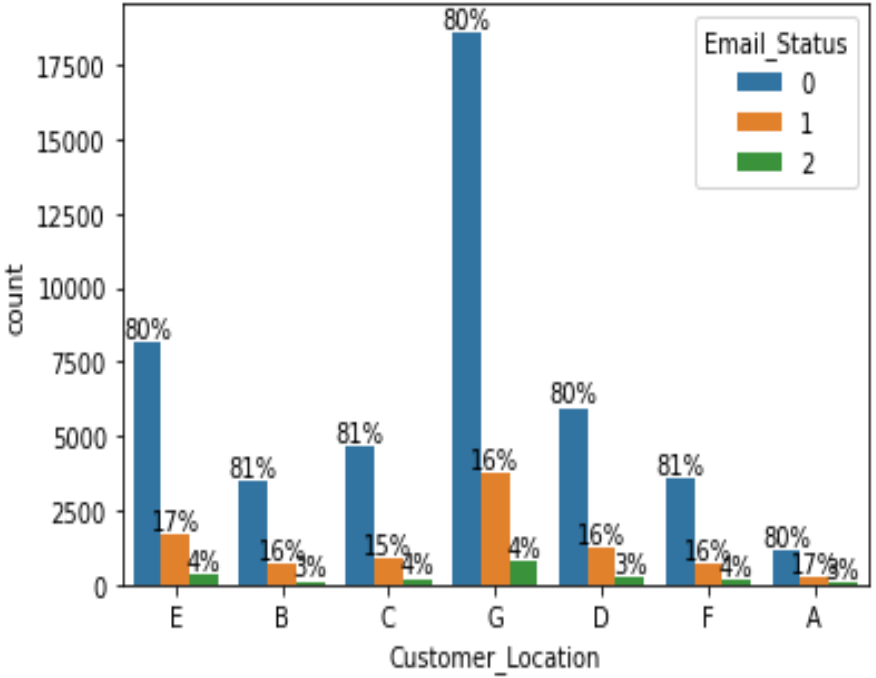
Email Sent by this categories



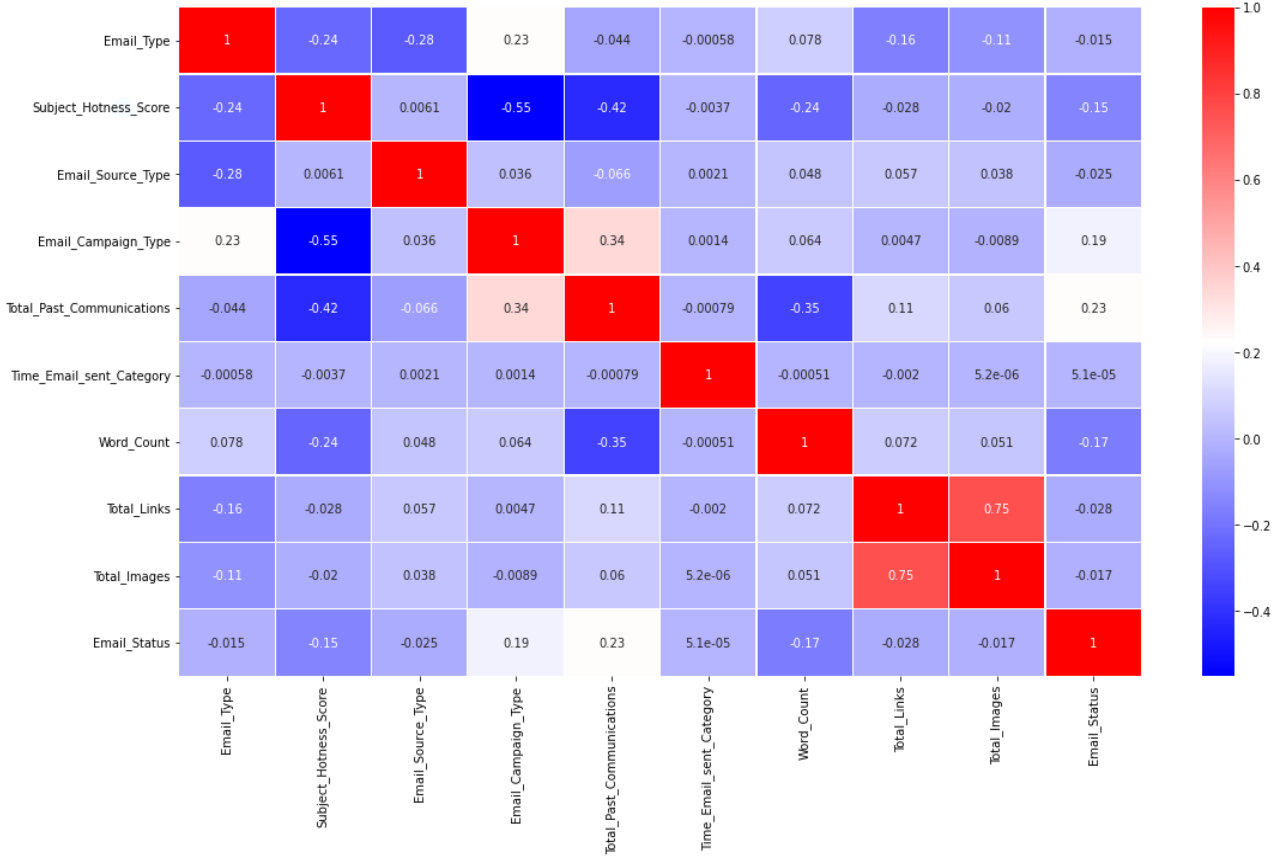


EDA (continued)





CORRELATION



1) Email_status is correlated with Total Past Communications, and minorly correlated to others

2) Email Campaign Type is highly correlated with Total Past Communications

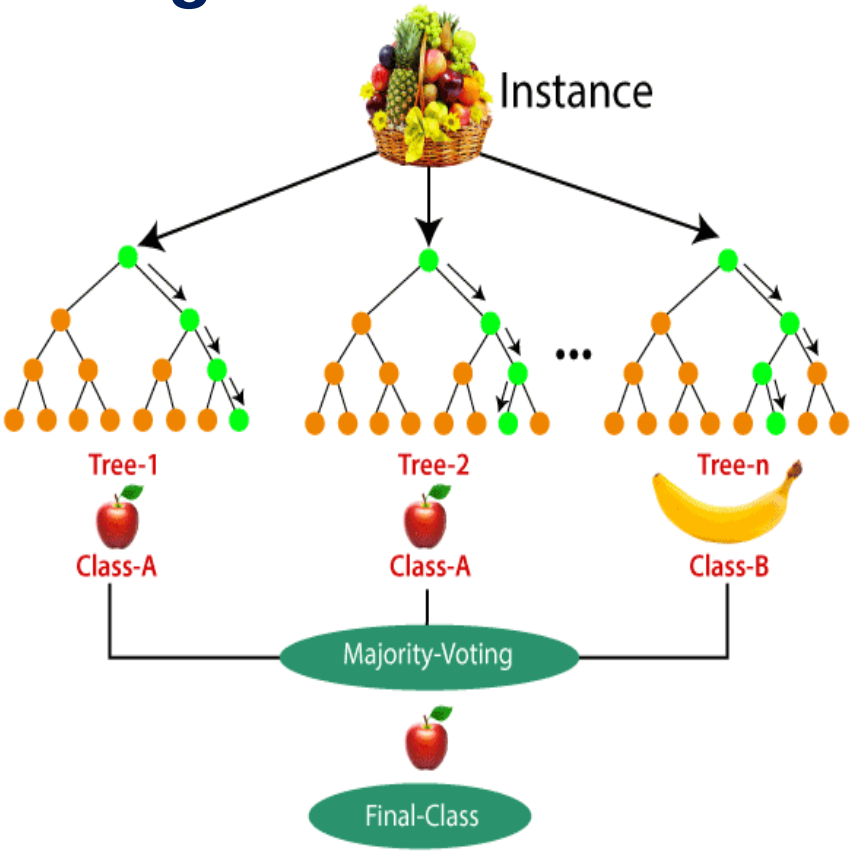
Machine Learning Model Building

- **Random Forests:**

- Random Forest Regression is an ensemble learning algorithm that operates by aggregating many random decision trees to make predictions while avoiding overfitting.

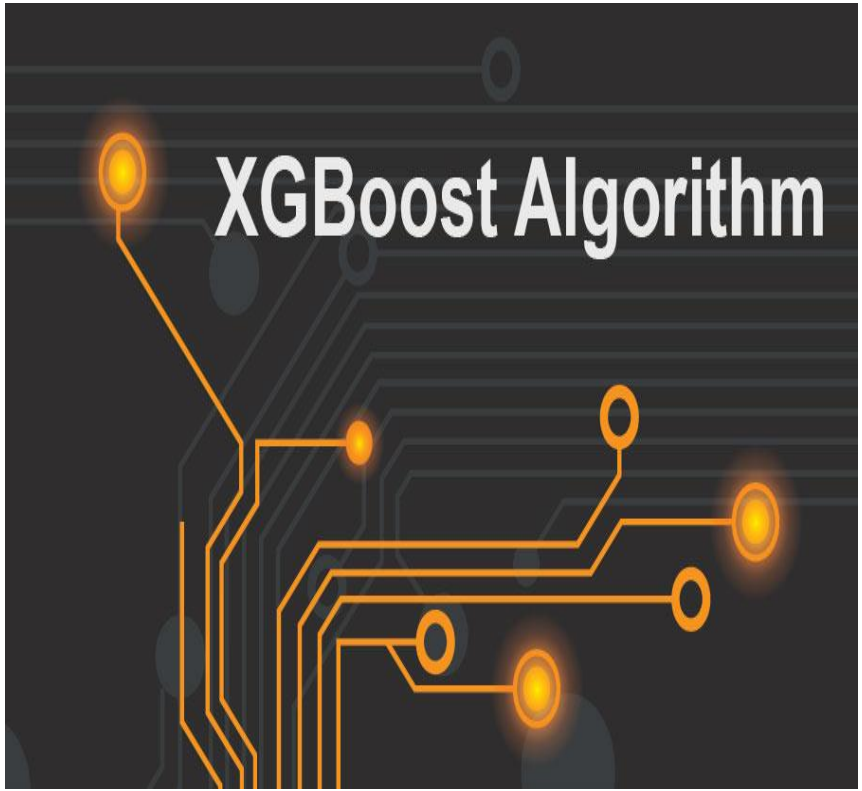
- **Advantages**

- 1) More accurate than others
- 2) Robust
- 3) Doesn’t face the overfitting
- 4) Highly versatile algorithm



XGBoost:

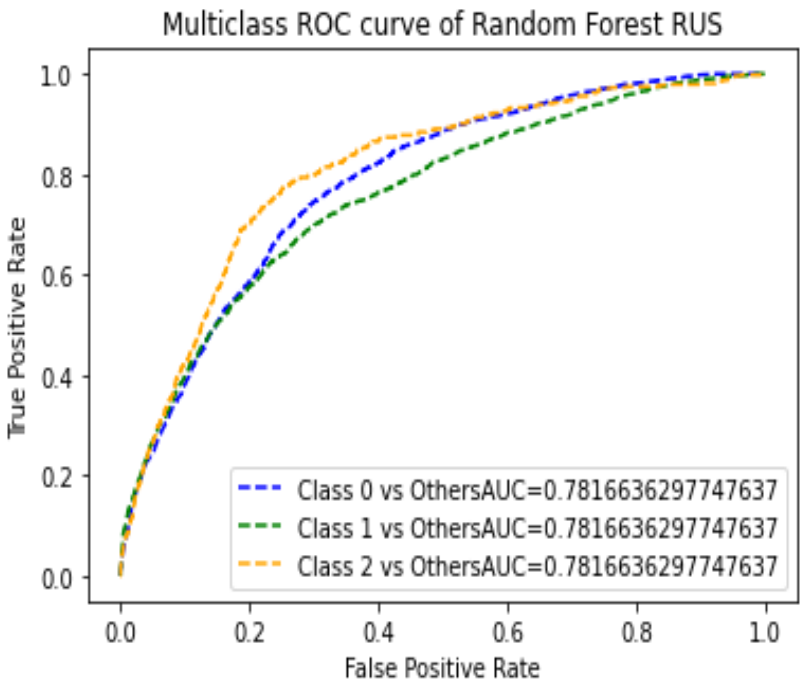
XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

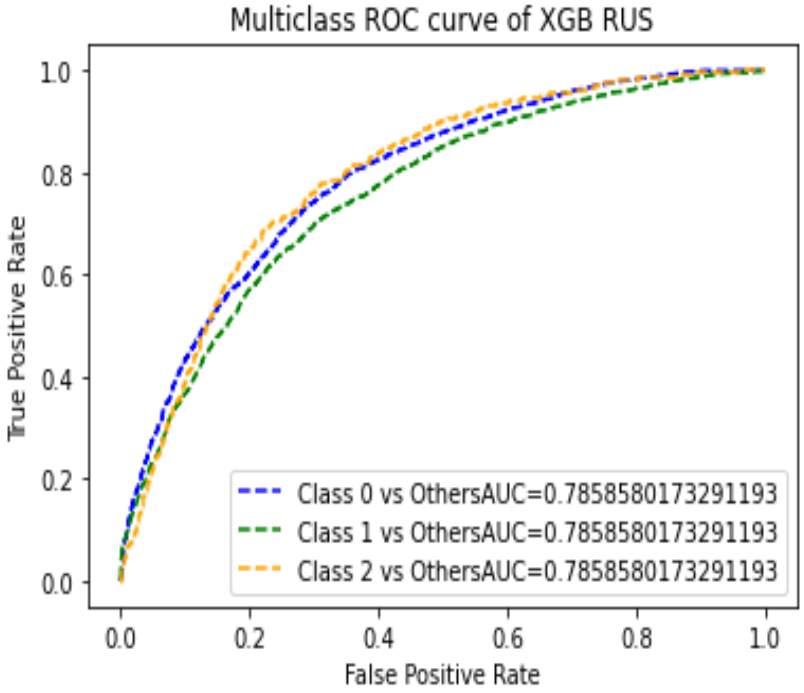


Advantages

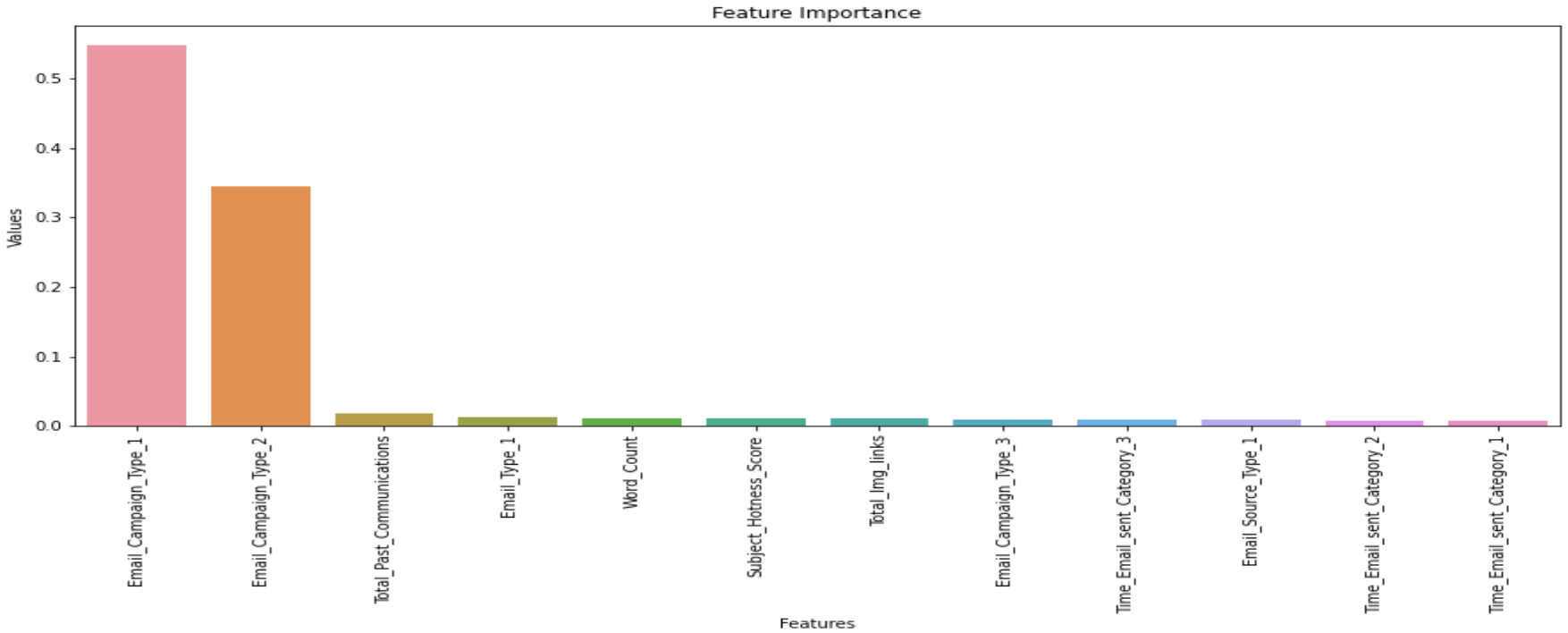
1) More accurate than random forest
2) Very good Performance

High Performance Model Results





Feature Importance



CONCLUSION

- 1) 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 the same irrespective of the demographic location or the time frame the emails were sent on.

- 2) In the Email Campaign Type feature, it seems like in campaign type 1 very few emails were sent but has a very high likelihood of getting read. Most emails were sent under email campaign type 2 and most ignored. Seems like campaign 3 was a success as even when less number of emails were sent under campaign 3, more emails were read and acknowledged.

- 3) Analyzing total past communications, we can see that the more the number of previous emails, the more it leads to read and acknowledged emails. This is just about making connection with your customers.

- 4) The more the words in an email, the more it has a tendency it has to get ignored. Too lengthy emails are getting ignored.

5) More images were there in ignored emails.

6) There are outliers in almost every continuous variable except Word Count and upon analyzing, it was found that outliers make up for more than 5% of the minority data and will influence the results either way, so it was better not to get rid of them.

7) There are outliers in almost every continuous variable except Word Count and upon analyzing, it was found that outliers make up for more than 5% of the minority data and will influence the results either way, so it was better not to get rid of them.

8)Email Campaign Type 1 and 2 are doing better than 3. So, focusing on improving 3, can do the tri ck.

9)The word count should be reasonable. The content should be crisp and to the point with a few m arketing gimmicks.

10)Total past communications had a positive influence, hence having a healthy relationship with cu stomers is a big yes.

THANK YOU