2024 Data Analysis

Evaluating Email Outreach Performance with Advanced Machine Learning Techniques

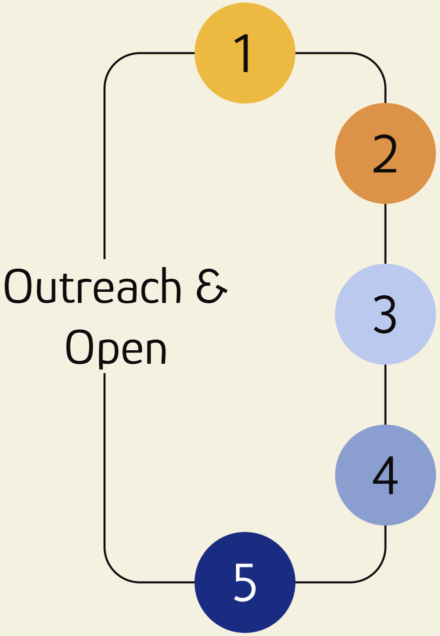
Riya, Sanika,
Vaishnavi, Aswad

Table of Contents

Project Description

- Background:
  - This project analyzes email outreach data to improve campaign effectiveness by understanding effect of the time email sent, revenue of the company, the date, company type and number of employees, and location of the company.
- Problem Statement:
  - Identify factors influencing the likelihood of recipients engaging with the email (opens) and increase the outreach efficiency.
  - Dependent Variable (DV): Email open status (opened or not).
  - Independent Variables (IVs): City - Home, Date Sent, Time Sent, State, Company type, Revenue, Headcount.

Data Preprocessing



Step 1	Dropped Unnecessary Columns: 'Bounces', 'Bounce Rate', 'Unsubscribes', and 'Unsubscribe Rate'
Step 2	Adding a New Column: 'Opened At' with default 1 to emails opened
Step 3	Removing Rows with 'Unsubscribed' status
Step 4	Checking for Duplicates and removing them
Step 5	Dropping Irrelevant Columns: 'Email status', 'Email permission status', 'Confirmed Opt-Out Date', 'Confirmed Opt-Out Reason', 'Website', 'Title', 'Country - Home'

Data Preprocessing

Sent & Merge: Open_Sent

1	Step 1	Removed Unsubscribed Records, Remove Duplicates: 'Campaign Name', 'Company', 'City - Home', and 'Created At'
2	Step 2	Dropping Irrelevant Columns just like Open dataset
3	Step 3	Left join Sent and Open datasets using 'Campaign Name' and 'Company' columns, Dropping Duplicate Columns with suffixes _x for 'sent' and _y for 'open'
4	Step 4	Renaming columns like 'City - Home_x': 'City - Home', 'State/Province - Home_x': 'State/Province - Home', 'Created At_x': 'Created At', 'Updated At_x': 'Updated At'
5	Step 5	Filled missing values in 'Opened At' with 0 and ensure there are no duplicates in Open_Sent which is merged with Outreach summary dataset using ‘Campaign Name’

Data Preprocessing

Email_outreach

1 2 3 4

Step 1

Converted 'Sent At' to date and time columns and drop irrelevant columns like 'Open Rate' and 'Click Rate' in Email_outreach

Step 2

State Mapping to standardize the names of states and create a new ‘State’ column

Step 3

Remove rows with most null values in duplicated rows

Step 4

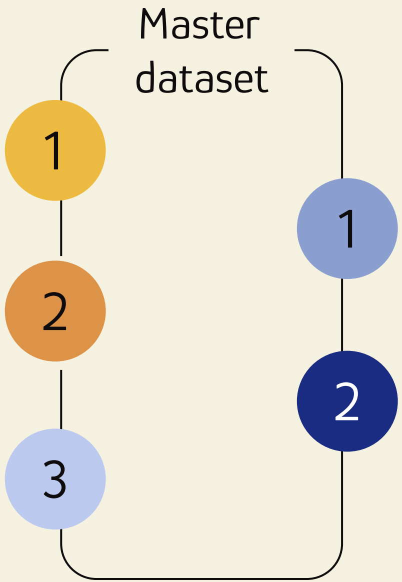
Remove any further duplicates based on Company, Campaign Name, and Time_Sent while keeping the first occurrence

Data Preprocessing

1 Company names from various sources were normalized by converting them to lowercase and removing special characters to ensure consistency.

2 Missing city, state, revenue, headcount, and other key columns were populated using the mode into new columns.

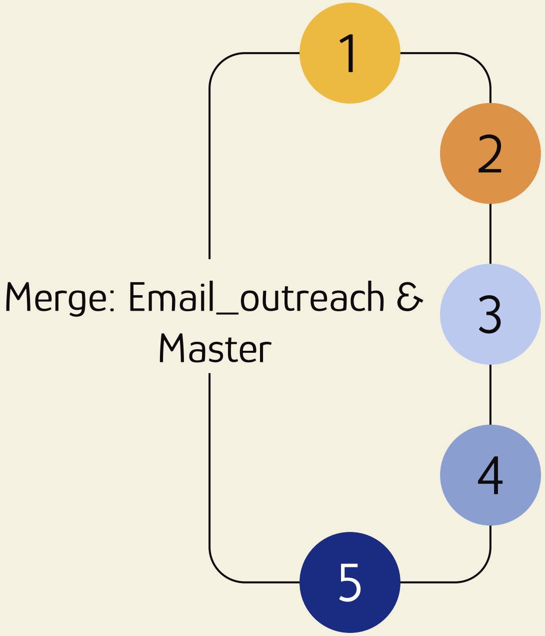
3 Duplicate rows based on the Company column were removed and unnecessary columns were removed.



1 After normalization, stemming was applied to company names using the Porter Stemmer to reduce words to their root forms (e.g., "advanced" to "advanc").

2 The TF-IDF vectorizer was used to transform the stemmed company names into numerical vectors. Cosine similarity scores were calculated between Email_outreach and master datasets to identify the most likely matches based on the similarity of their company names.

Data Preprocessing



Step 1 — Company name in Email_outreach is normalized, and the first word is extracted and this is matched with master dataset columns

Step 2 — Calculate the most frequent (mode) value for each column related to city, state, revenue, headcount, and zipcode in master

Step 3 — Inner join on Email_outreach and Master datasets on ‘Company’ column

Step 4 — State and City-Home columns from Master are used to fill the null values in State and City of Email_outreach dataset

Step 5 — Redundant columns like State_master, City - Home_master are removed

Final dataset: Merged1

01 Campaign Name

02 Company

03 City - Home

04 Date Sent

05 State

06 Revenue

07 Opened

Binary Independent Variable

08 Mobile open rate

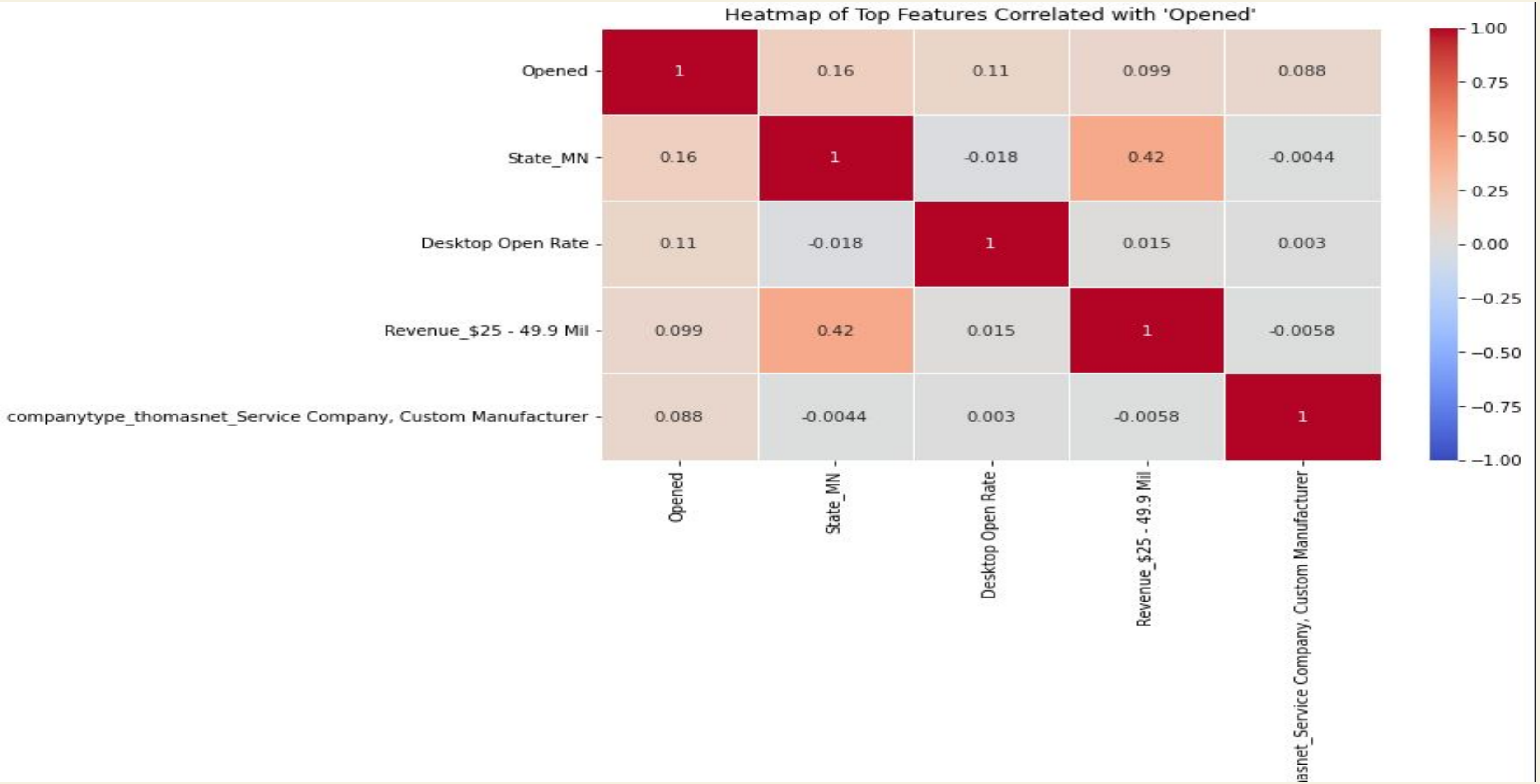
09 Desktop open rate

10 Time_Sent

11 Company_type

12 Headcount

Correlation Matrix



Data Analysis : Training the model

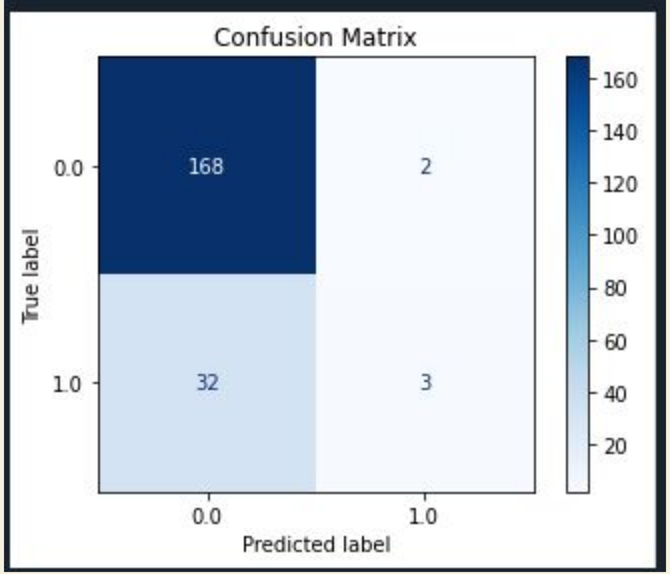
Random Forest Model: It is a learning method used for both classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

Our Findings:

Accuracy: 0.8341463414634146

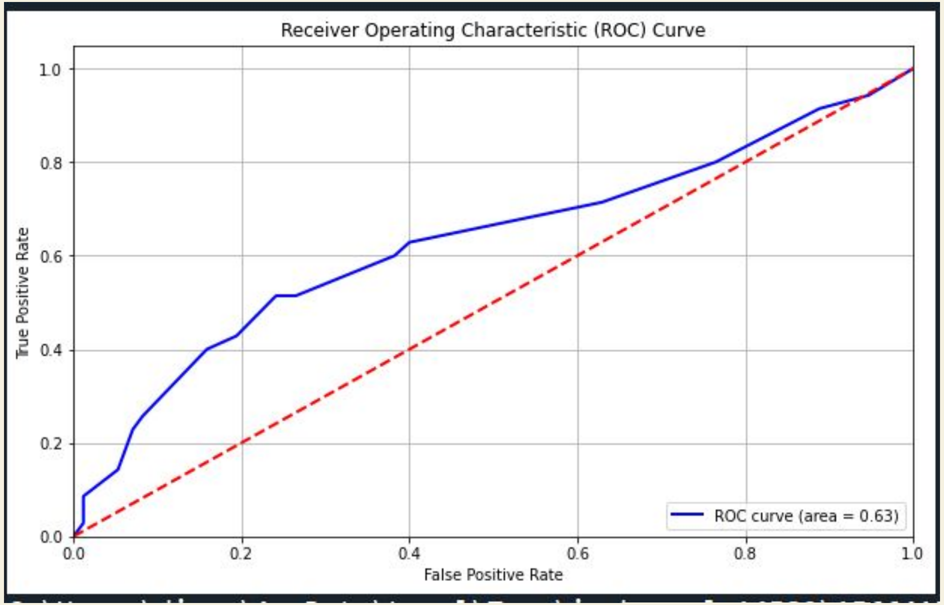
ROC-AUC Score: 0.6298319327731092

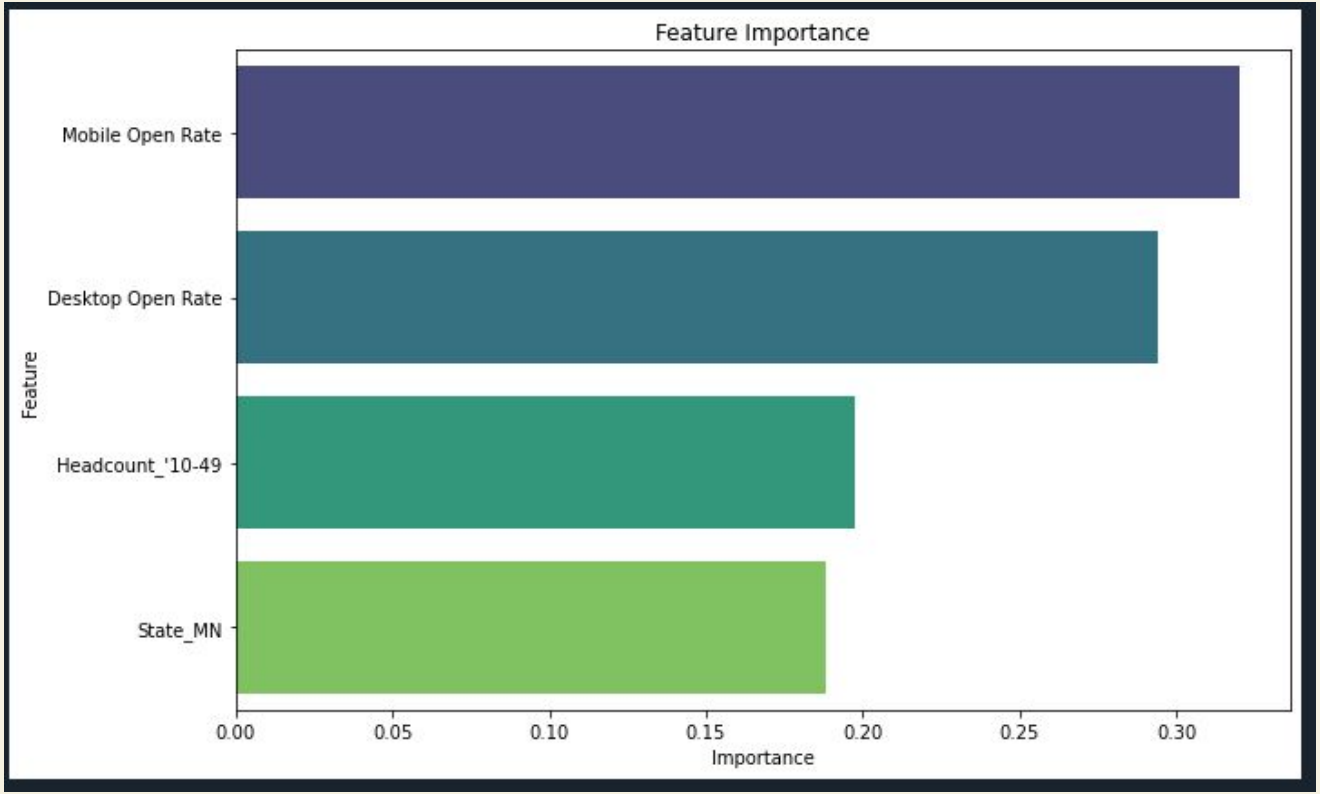
	precision	recall	f1-score	support
0.0	0.84	0.99	0.91	170
1.0	0.60	0.09	0.15	35
accuracy			0.83	205
macro avg	0.72	0.54	0.53	205
weighted avg	0.80	0.83	0.78	205



Confusion Matrix

Receiver Operating Characteristic (ROC) Curve

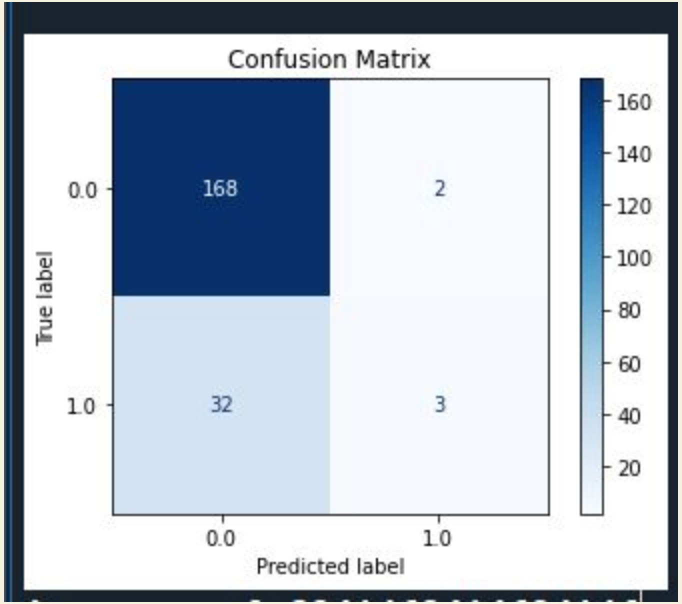




Feature Importance

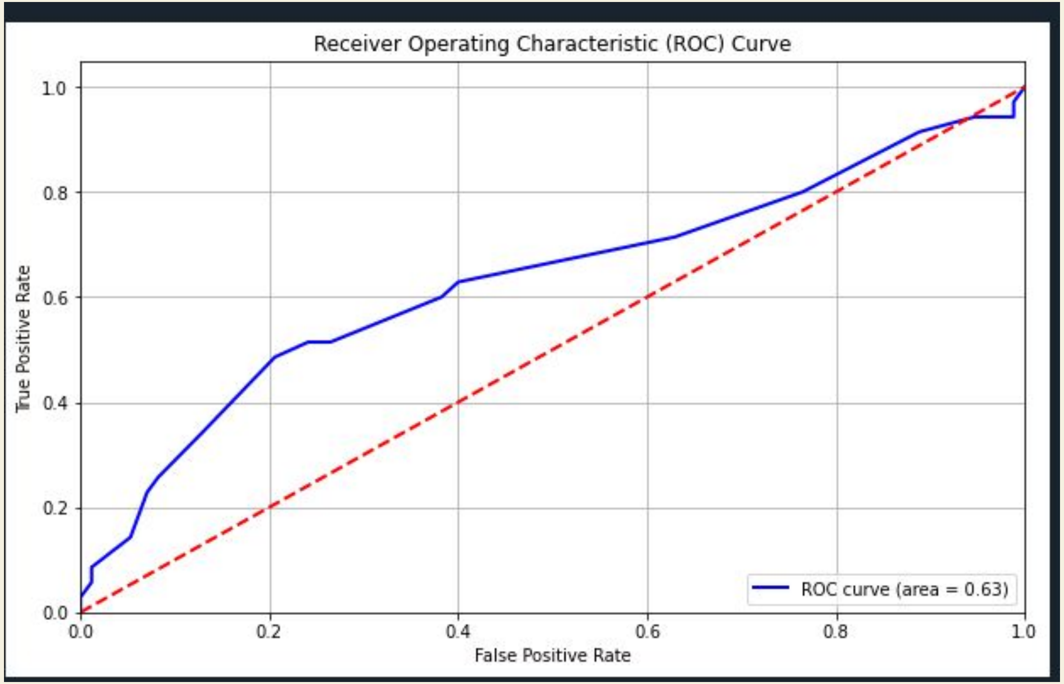
Gradient Boosting Machine(GBM): It is an ensemble learning algorithm used for classification tasks. It builds a strong predictive model by combining the predictions of multiple weaker models, typically decision trees, in a sequential manner.
Our Findings:
Accuracy: 0.8341463414634146
ROC-AUC Score: 0.6306722689075631

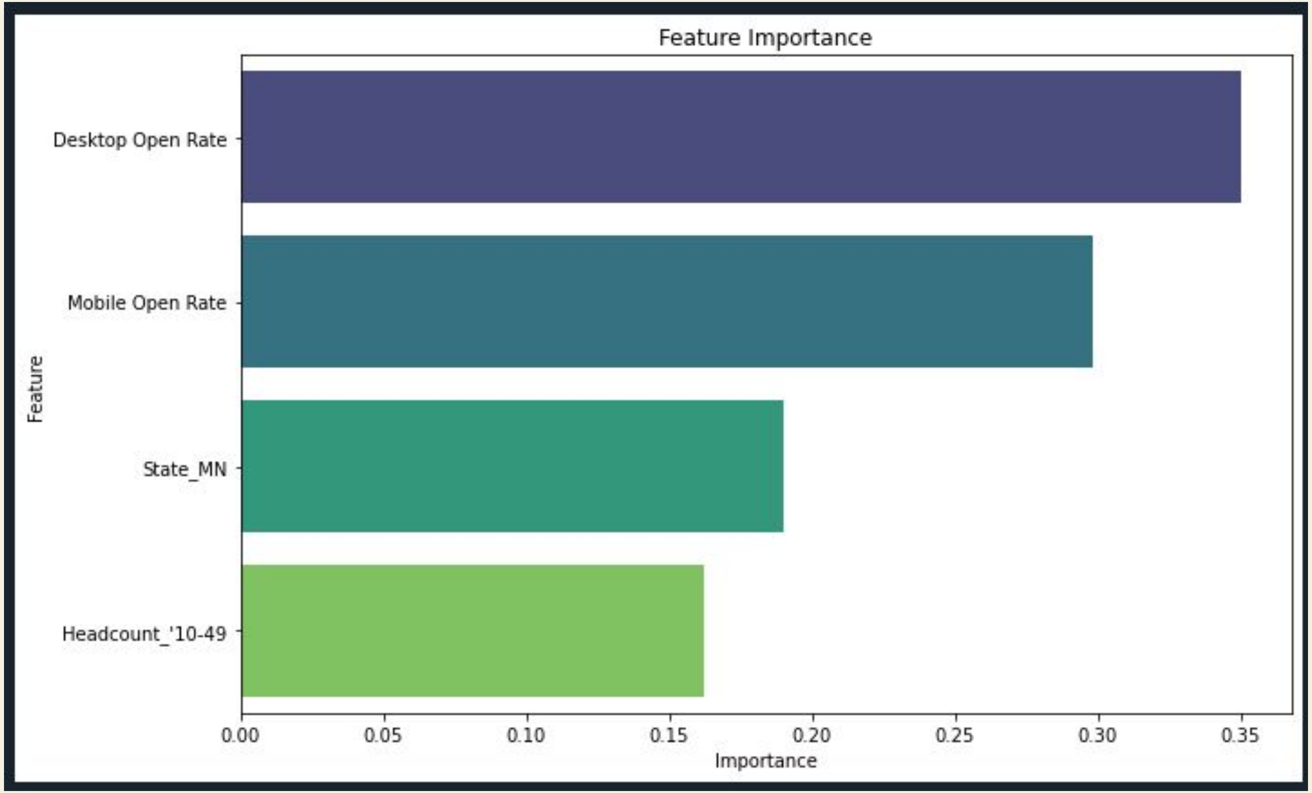
	precision	recall	f1-score	support
0.0	0.84	0.99	0.91	170
1.0	0.60	0.09	0.15	35
accuracy			0.83	205
macro avg	0.72	0.54	0.53	205
weighted avg	0.80	0.83	0.78	205



Confusion Matrix

Receiver Operating Characteristic (ROC) curve





Feature Importance

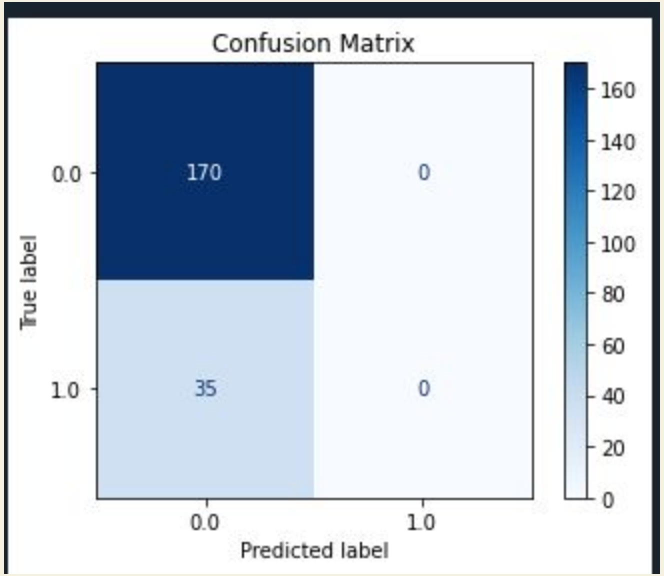
Support Vector Machine (SVM) a machine learning algorithm that classifies data and solves regression tasks. SVMs are particularly good at binary classification problems, where data is separated into two groups.

Our Findings:

Accuracy: 0.8292682926829268

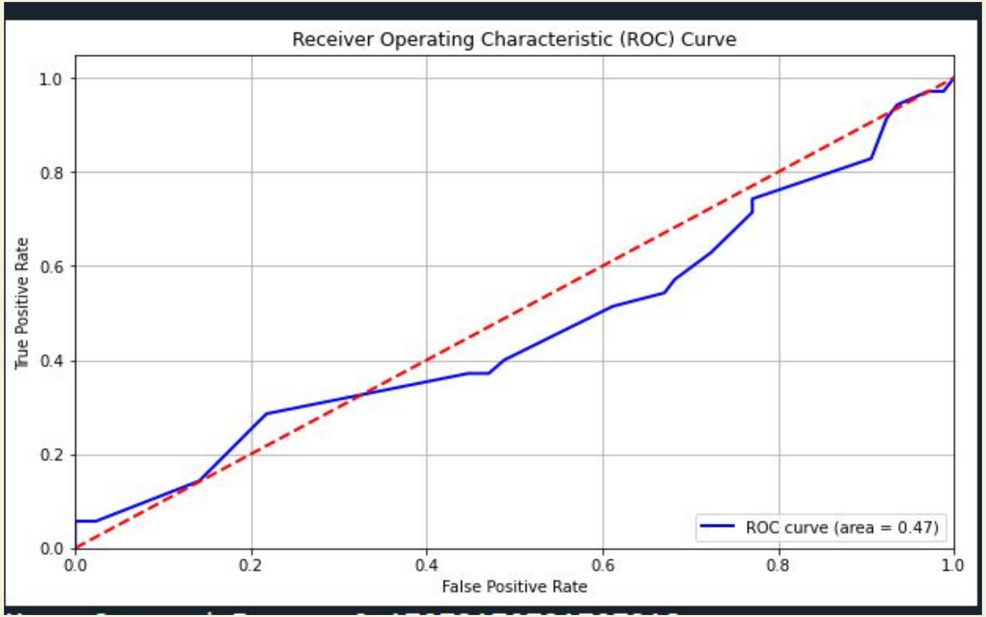
ROC-AUC Score: 0.4659663865546218

	precision	recall	f1-score	support
0.0	0.83	1	0.91	170
1.0	0.00	0.00	0.00	35
accuracy			0.83	205
macro avg	0.41	0.50	0.45	205
weighted avg	0.69	0.83	0.75	205



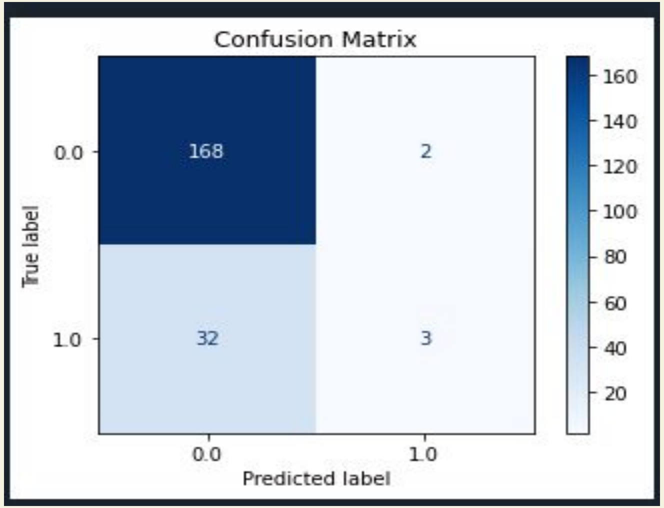
Confusion Matrix

Receiver Operating Characteristic (ROC) Curve



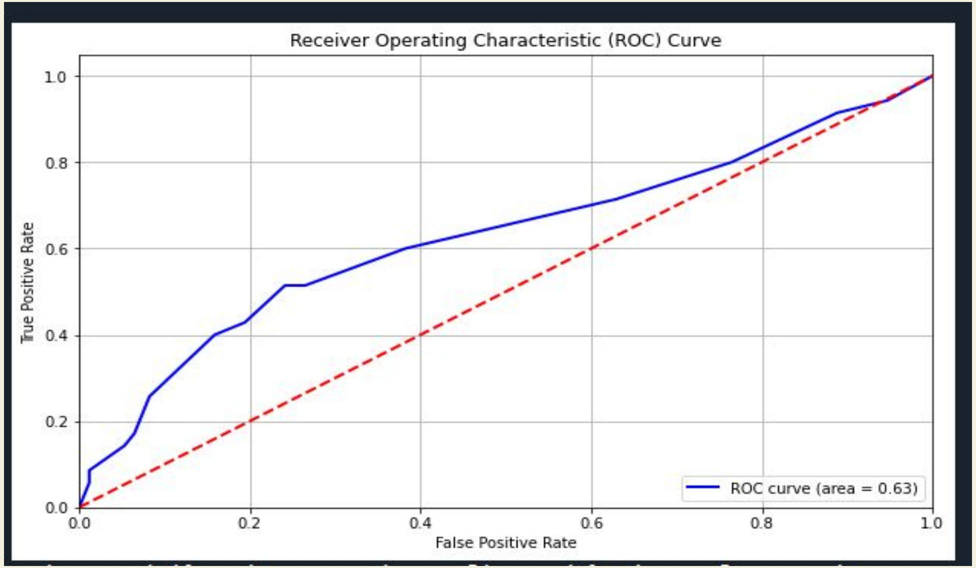
Decision Tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks.
Our Findings:
Accuracy: 0.8341463414634146
ROC-AUC Score: 0.6269747899159664

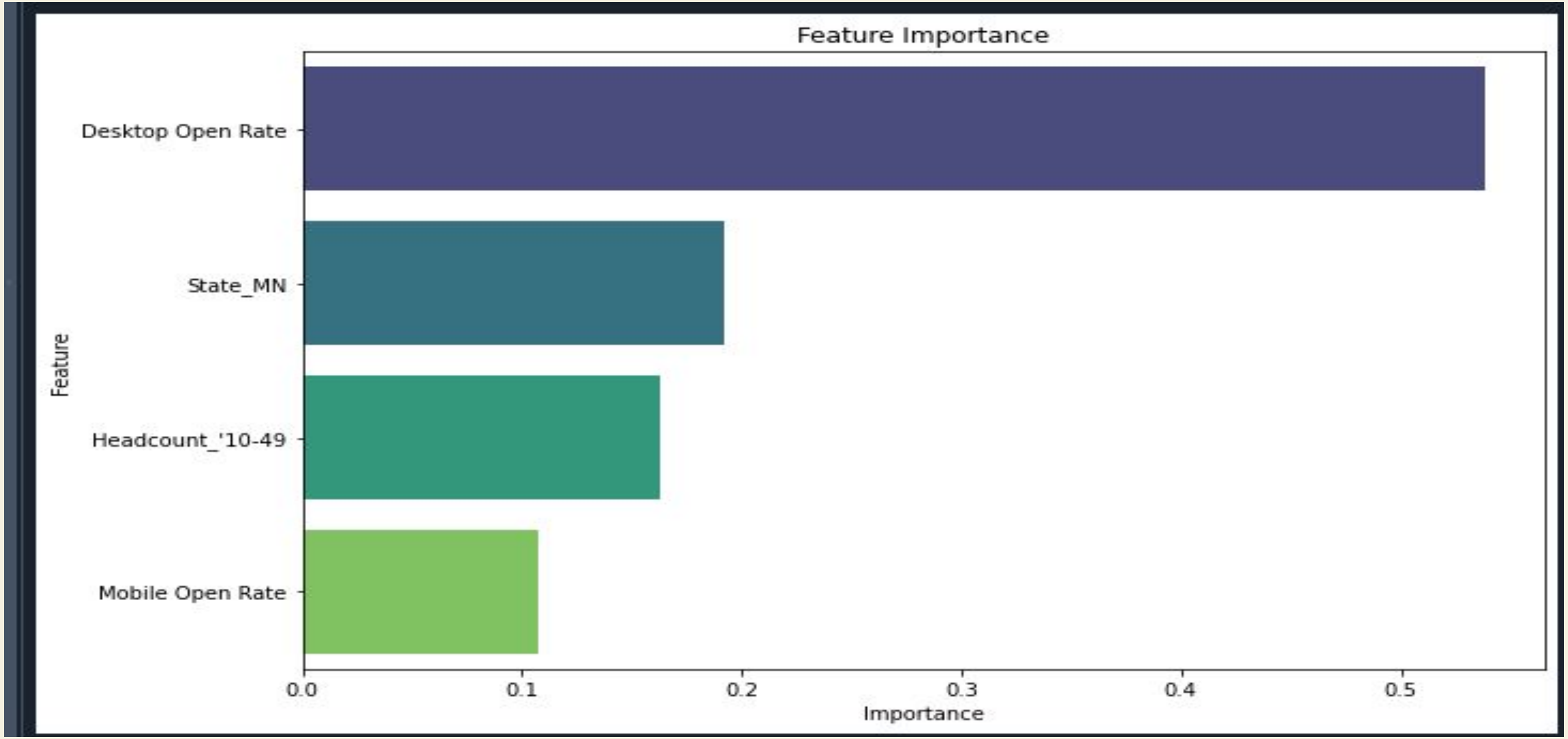
	precision	recall	f1-score	support
0.0	0.84	0.99	0.91	170
1.0	0.60	0.09	0.15	35
accuracy			0.83	205
macro avg	0.72	0.54	0.53	205
weighted avg	0.80	0.93	0.78	205



Confusion Matrix

Receiver Operating Characteristic (ROC) Curve





Feature Importance

Results

Random Forest	Gradient Boosting Machine	Support Vector Machine(SVM)	Decision Tree
Accuracy: 0.8341 **ROC-AUC Score**: 0.6298	**Accuracy**: 0.8341 **ROC-AUC Score**: 0.6307	**Accuracy**: 0.8293 **ROC-AUC Score**: 0.4660	**Accuracy**: 0.8341 **ROC-AUC Score**: 0.6270

- ❑ Since, correlation was strong between Desktop Open Rate, State_MN, Headcount '10-49 and Mobile Open Rate and Opened, we chose them for our independent variable.
- ❑ We found out that after comparing all the models, Training the company data with Gradient Boosting Machine (GBM) is the best fit as it has the more ROU - AUC score in compare to other models.

Suggestions / Recommendations

- Final dataset can be increased by matching companies using other methods like Stemming, or other algorithms.
  - Stemming was tried which reduced number of rows to 706.
- Consistent way of recording data can make this analysis more focussed.
- Follow-up emails could be targeted more effectively to non-openers to increase interaction.
- Content of the email: Subject and going to spam can also impact the results of this analysis.

Summary of the Group Project

Biggest Challenge

Merging datasets especially master

Aspect Needing Improvement

The merging process could be further refined

Most Well-Done Aspect

Merging and modeling

Reference

- GeeksforGeeks: Provided guidance on data preprocessing, merging datasets, and applying specific Python functions.
  - Website: https://www.geeksforgeeks.org
- W3Schools: Helped with syntax, function usage, and understanding pandas, regex, and other Python-related concepts.
  - Website: https://www.w3schools.com
- For training the models like SVM, Decision tree and others:
  - Website: https://scikit-learn.org/stable/modules/svm.html
  - https://scikit-learn.org/stable/modules/tree.html
  - https://scikit-learn.org/stable/modules/ensemble.html

Questions

Task Division and Communication

Data Preprocessing: Outreach, Sent, and Open datasets	Handled by Riya and Vaishnavi.
Data Preprocessing: Master dataset	Handled by Sanika, Riya, Aswad and Vaishnavi.
Merging	Collaborative effort by Sanika, Riya, and Vaishnavi.
Modeling	Managed by Riya and Sanika.
Presentation	Created by Vaishnavi and Aswad.
Communication	Whatsapp, Google Colab, Google docs, Gmail