Email Campaign Analysis

Dataset Overview

- **Subject**: Represents the email subject.
- **Body**: Content of the email.
- **Opened**: Indicates if an email was opened (assumed).
- **Meeting Link Clicked**: Shows if a meeting link was clicked (assumed).
- **Responded**: Reflects if a user responded to the email (assumed).

Data Processing

- In the dataset which consists of a lists of dictionaries, a single term 'HR_Consulting_Series' is observed which is removed to maintain data consistency.
- After transforming the pickled data into a dataframe, there were two columns created: "meeting link clicked" and "meeting_link_cleaned." These columns were merged to form a unified column named “meeting link cleaned.”
- Binary categorical columns were filled with values of 1 and 0 to denote their respective categories in the dataframe.

Exploratory Data Analysis (EDA)

Summary:

- Total number of instances are 154 which have no null values.
- 46 out of 154 users have clicked the meeting.
- 8 out of 154 users responded.
- 114 out of 154 have opened.

Analysis:

- Approximately 74% of the users are opening the received emails, we may account for the good subject of the email or users having interest to view.
- But only 29.8% users are clicking on the meeting and 5.1% responding to it. The response rate for the opened emails is 7.0%.
- Since users clicking the meeting is less compared to users opening it, further analysis was done on it.

Word Cloud Insights

- For emails where the meeting link was clicked, words like 'analytics', 'strategy', and 'solutions' stand out. This could suggest that recipients are looking for actionable and analytical insights that can help them in decision-making.
- In contrast, for emails that did not result in a click, terms like 'marketing', 'data', and 'insights' were prevalent. It's possible these terms on their own are becoming too generic and not as compelling without a clear actionable context.

Sentiment Analysis Insights

- The average sentiment score for emails where the meeting link was clicked ('meeting link clicked 1') is 0.2308, which is comparatively lower than the score for emails without a click ('meeting link clicked 0') at 0.6019. This is counterintuitive, as one might expect a more positive sentiment to correlate with higher engagement.
- A potential explanation might be that a moderately positive or neutral tone in the email body is perceived as more professional and less aggressive, thereby encouraging more serious consideration and action like clicking on a meeting link.

Model Development and Analysis

Model Selection

- Models like logistic regression, random forest classifier, svm, gradient boosting, decision trees are used.
- These models range from foundational algorithms like Logistic Regression, which is straightforward and interpretable, to more complex ensemble methods like Gradient Boosting, which are powerful but less transparent.

Pipeline 1

- Converted words in subject and body into numerical using TF-IDF vectorization, this highlights the context of each word as well.
- Employed PCA with n=2, to reduce dimensionality.

Result Analysis

- The Logistic Regression and SVC models emerged as top performers with accuracy and F1 scores of approximately 87%.

- These models effectively balanced precision and recall, indicating a high rate of correct predictions for both 'opened' and 'not opened' classes.
- The initial findings suggest that a straightforward approach using Logistic Regression can provide strong predictive performance.
- This indicates that the patterns in the data are not overly complex, or that the most relevant information for predicting engagement is linearly separable.

Pipeline 2

- Extracted sentiment score of each body and subject using VADER and added them as features thereby increasing vector space.
  TF-IDF vectorization for body and subject columns.
- This captures words' context and users' emotion as well.

Result Analysis

- Post refinement, the RandomForest Classifier and SVC stood out with perfect recall scores, meaning they correctly identified all 'opened' emails in the test set.
- This suggests that sentiment can be a crucial factor in predicting email engagement.
- It can also suggest overfitting as data is imbalanced.

Insights and Actionable Insights

Insights

- **Open Rate versus Engagement:** The data shows a high open rate at 74%, indicating that subjects are effective in garnering initial interest. However, there is a significant drop when it comes to clicking on meeting links (29.8%) and responding (5.1%). This suggests that while subjects draw recipients in, the body content might not be as compelling in driving further action.
- **Influence of Specific Terms:** Words like 'analytics', 'strategy', and 'solutions' are more prevalent in emails where the meeting link was clicked. In contrast, 'marketing', 'data', and 'insights' appear often in emails that did not result in clicks, suggesting that certain terms may be overused and have lost their impact.
- **Sentiment Analysis Paradox:** The sentiment scores revealed an unexpected trend; emails that led to clicks had a lower sentiment score on

average than those that did not. This implies that a more neutral tone could be more conducive to driving engagement than overly positive language.

- **Model Performance:** Logistic Regression and SVC models performed best, suggesting that linear models are quite sufficient for this dataset. These models managed to effectively balance precision and recall, achieving high F1 scores.

Actionable Insights

- **Content Tailoring for Action:** Improving the email body to be more action centric.Instead of using generic terms, focus on clear and specific action words that have been shown to correlate with higher engagement, such as 'discover', 'achieve', and 'improve'.
- **Optimizing Email Tone:** Refraining from using overly positive or promotional language, which might be off-putting for some users. Instead, adopting a professional and moderately enthusiastic tone that aligns with the neutral to positive sentiment scores associated with higher click-through rates.
- **Personalization:** Tailoring emails based on users behavior, interests, and previous interactions.