# BookSends.com

## CTR Analytics

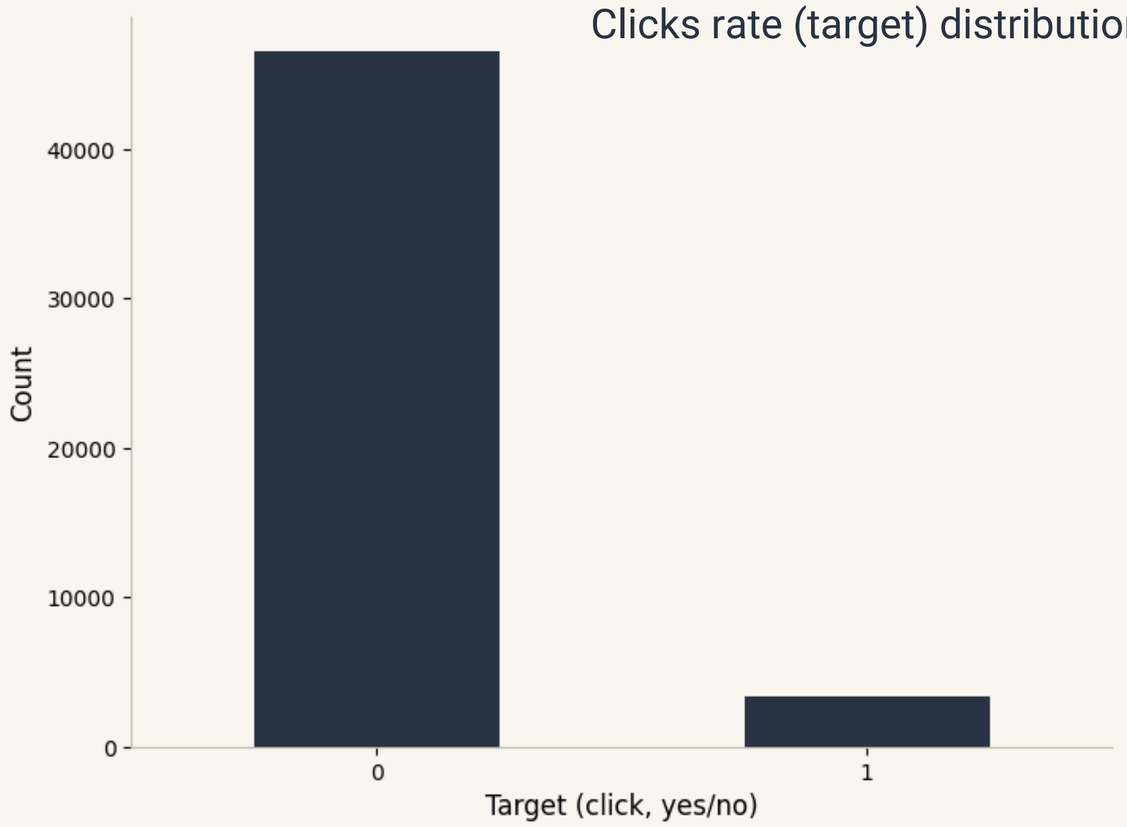
Elizaveta Saigina
MIB 3


# Ads and CTR

The online ad industry is a thriving digital marketplace, with an 8% annual growth rate projected to hit **$160B by 2023**. Despite consumers professing not to click on ads often, the internet's vast scale and the 'long tail' effect foster the industry's longevity and growth. Key to this vitality is the **Click-Through Rate (CTR)**, a critical metric for advertisers to gauge the success of their online campaigns. For BookSends.com, **maximizing CTR** is an essential business goal - and data science offers us the tools to achieve it. By predicting CTRs using historical email data, we can streamline ad placements, improve future emails, and provide robust data-driven advice on advertising rates.

# Dataset Analytics

**Insights:** Highly-unbalanced target, contains NaN and categorical features which require additional encoding.



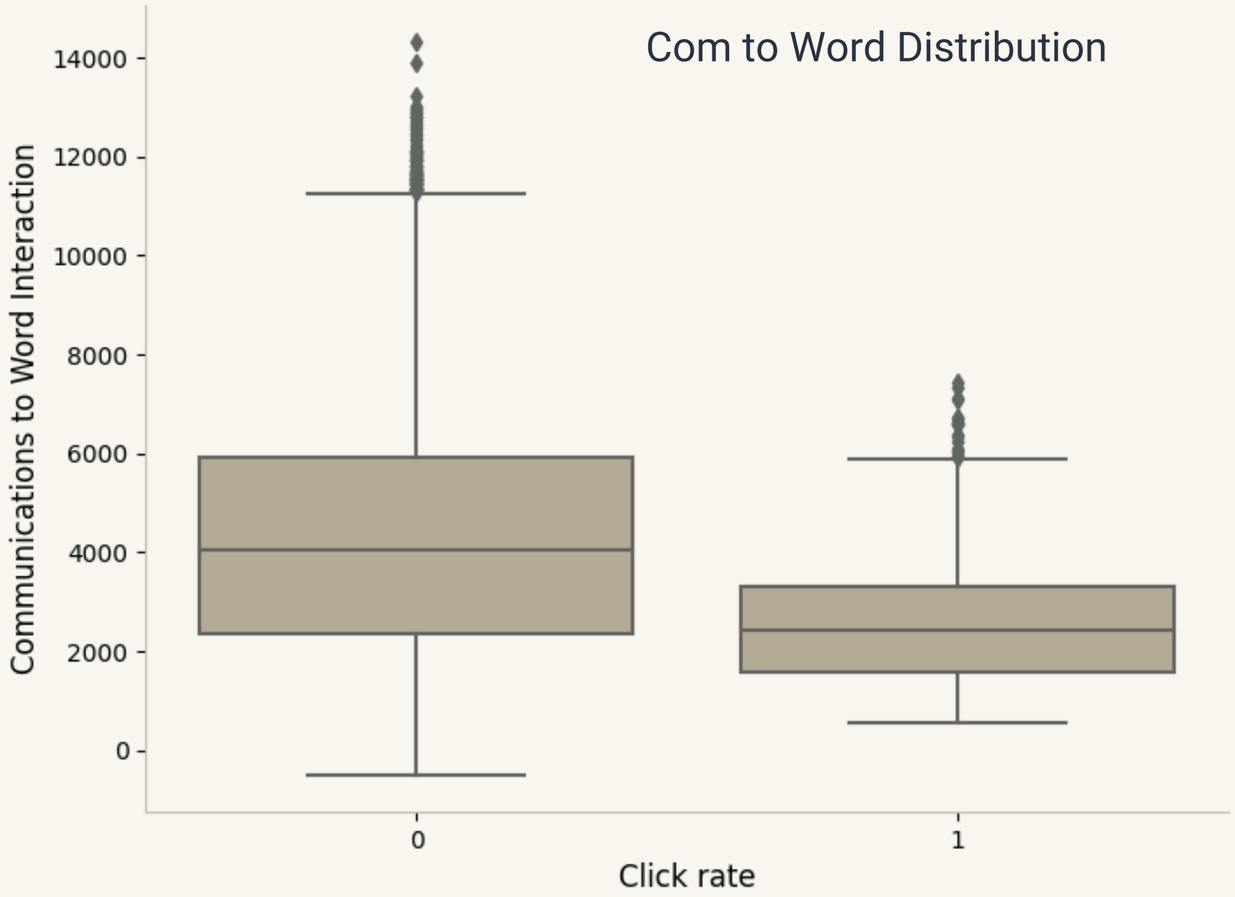

**50000**
Samples

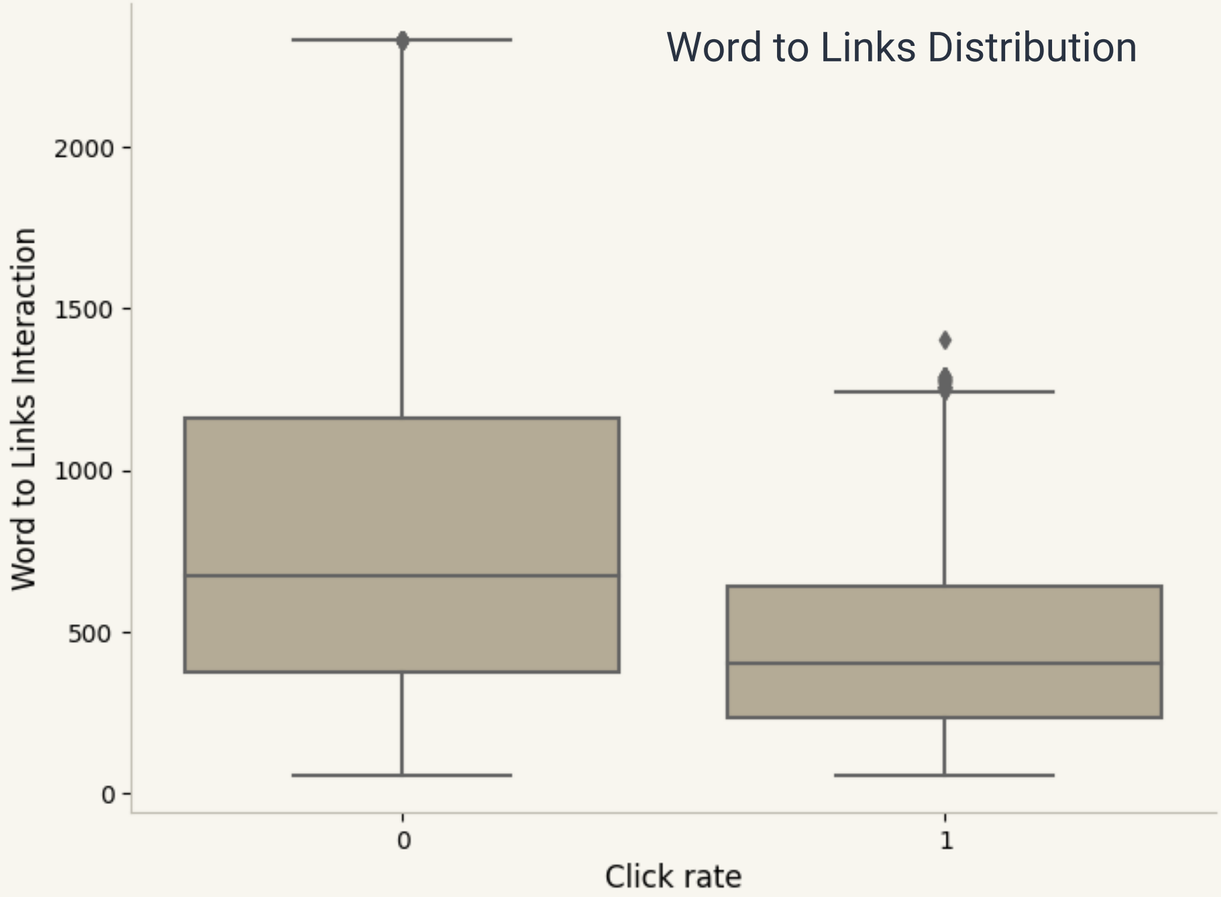
**55**
Features
**11** Initial and **44** created during EDA process

Features like *Sent_Hour* were transformed into categories like Day, Night, etc. and then encoded.

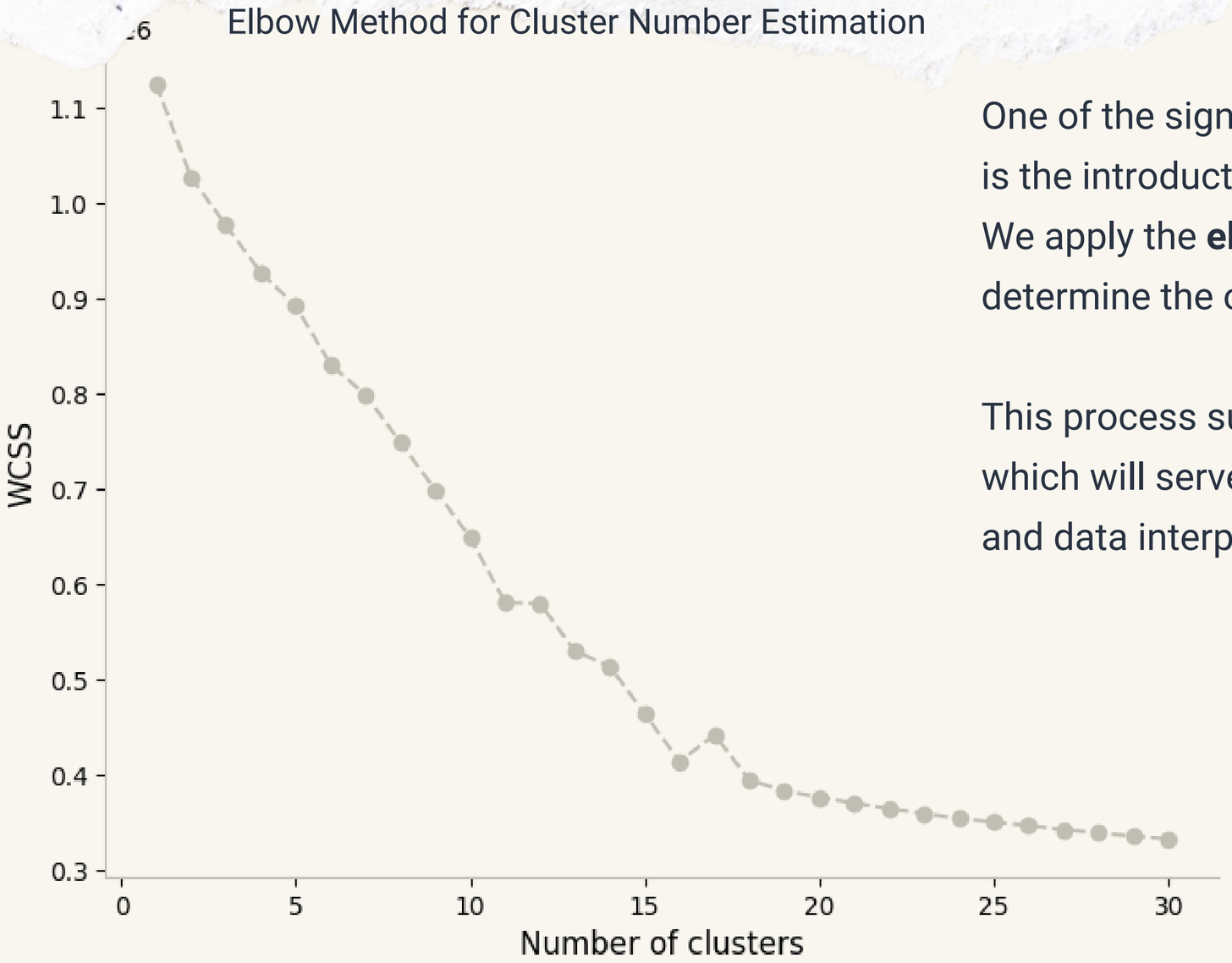
# Generating New Features

By marking missing values, multiplying columns, identifying outliers, and more, we created an enriched dataset that unveils deeper patterns. In particular, the multiplication of features has revealed significant insights.



# Hunting for Optimal Clusters

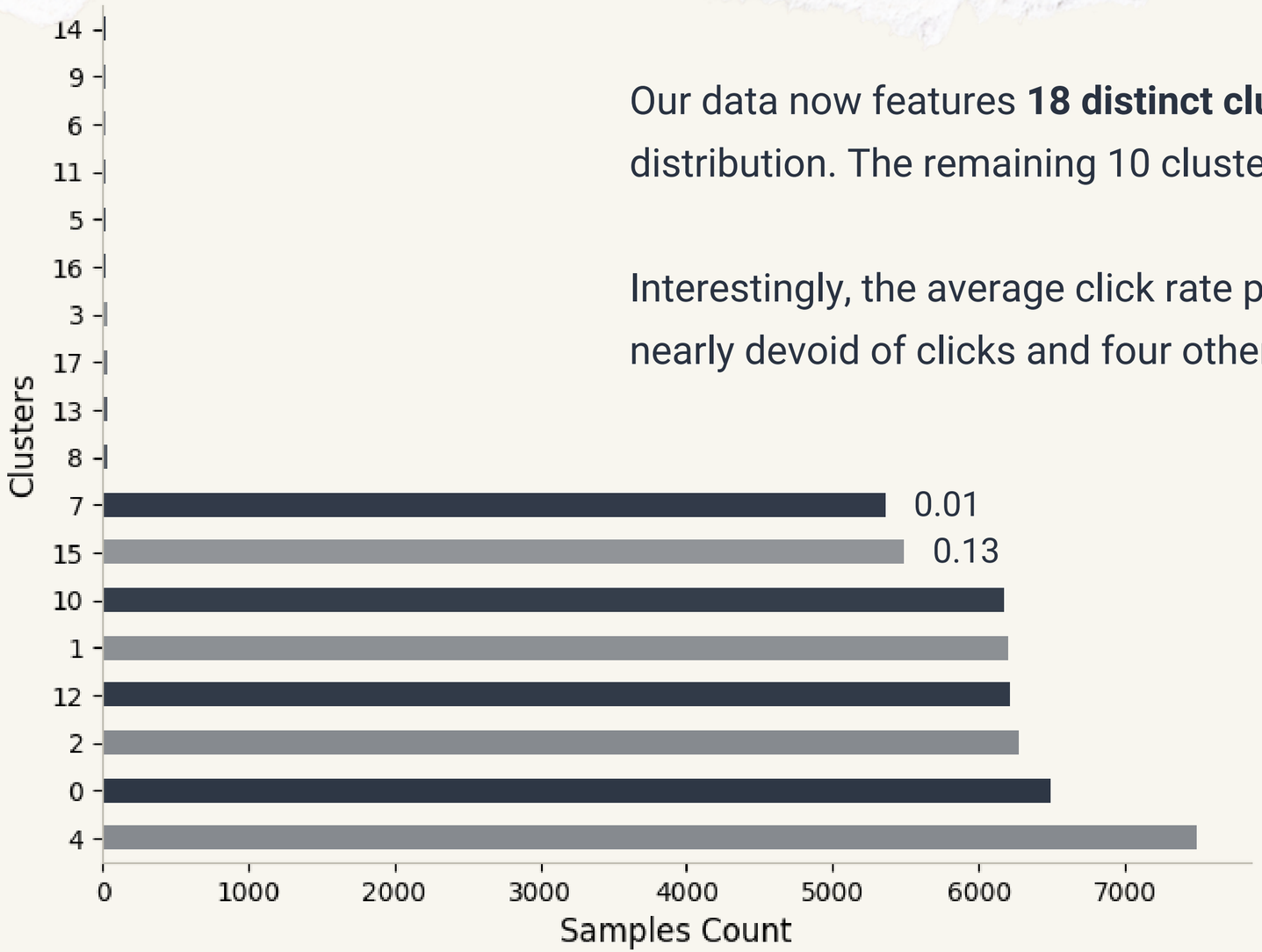



One of the significant enhancements made to our dataset is the introduction of clustering as an additional feature. We apply the **elbow rule** via a **K-means algorithm** to determine the optimal number of clusters.

This process suggested an optimal count of **18 clusters,** which will serve as a valuable feature for model training and data interpretation.

# Understanding the Clusters

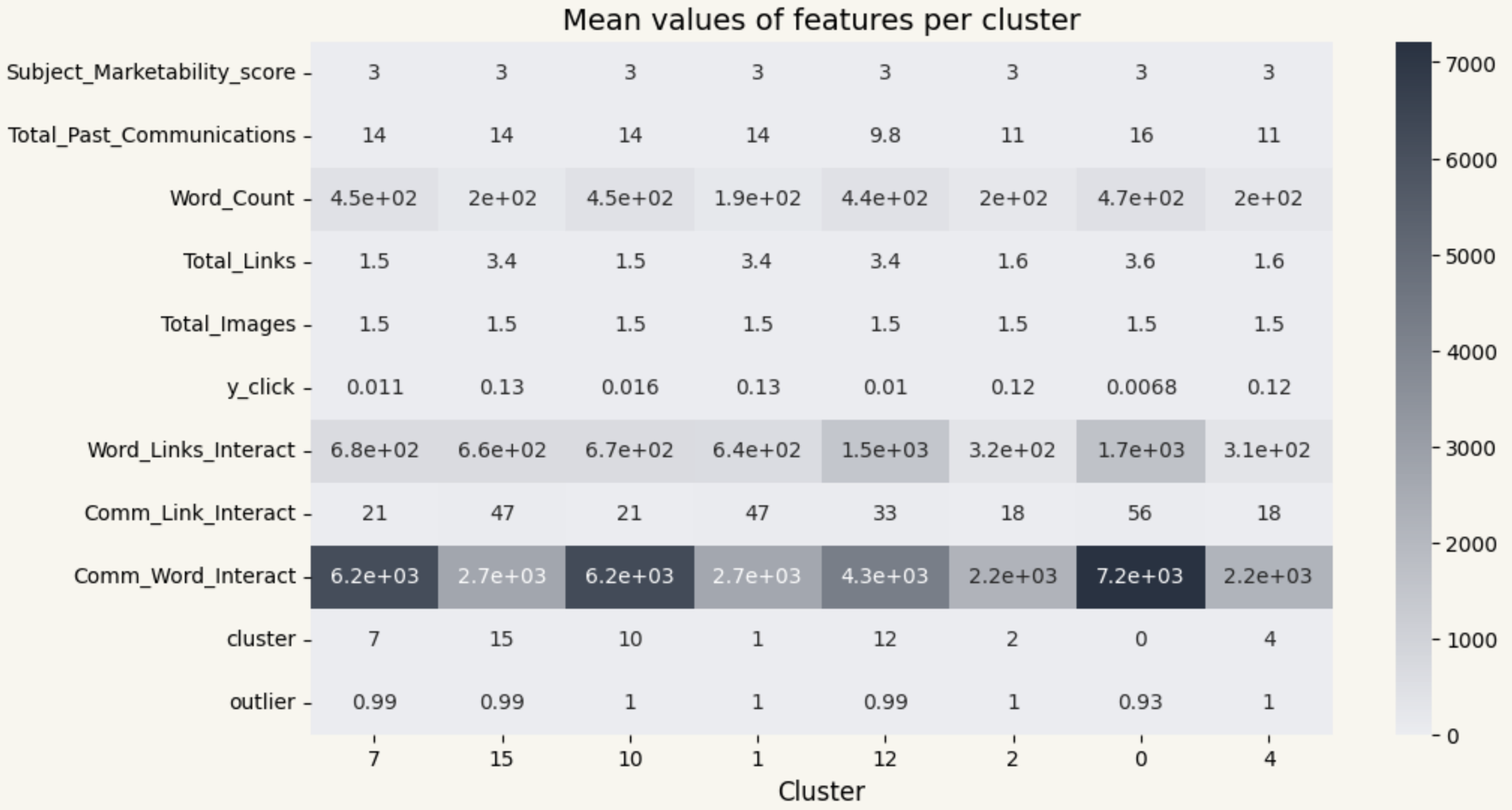
## A Detailed Look at Our Cluster Characteristics




Our data now features **18 distinct clusters**, with **8 main** ones dominating the distribution. The remaining 10 clusters contain less than 40 samples each.

Interestingly, the average click rate per cluster varies significantly, with four clusters nearly devoid of clicks and four others showing an average click rate of 0.13.

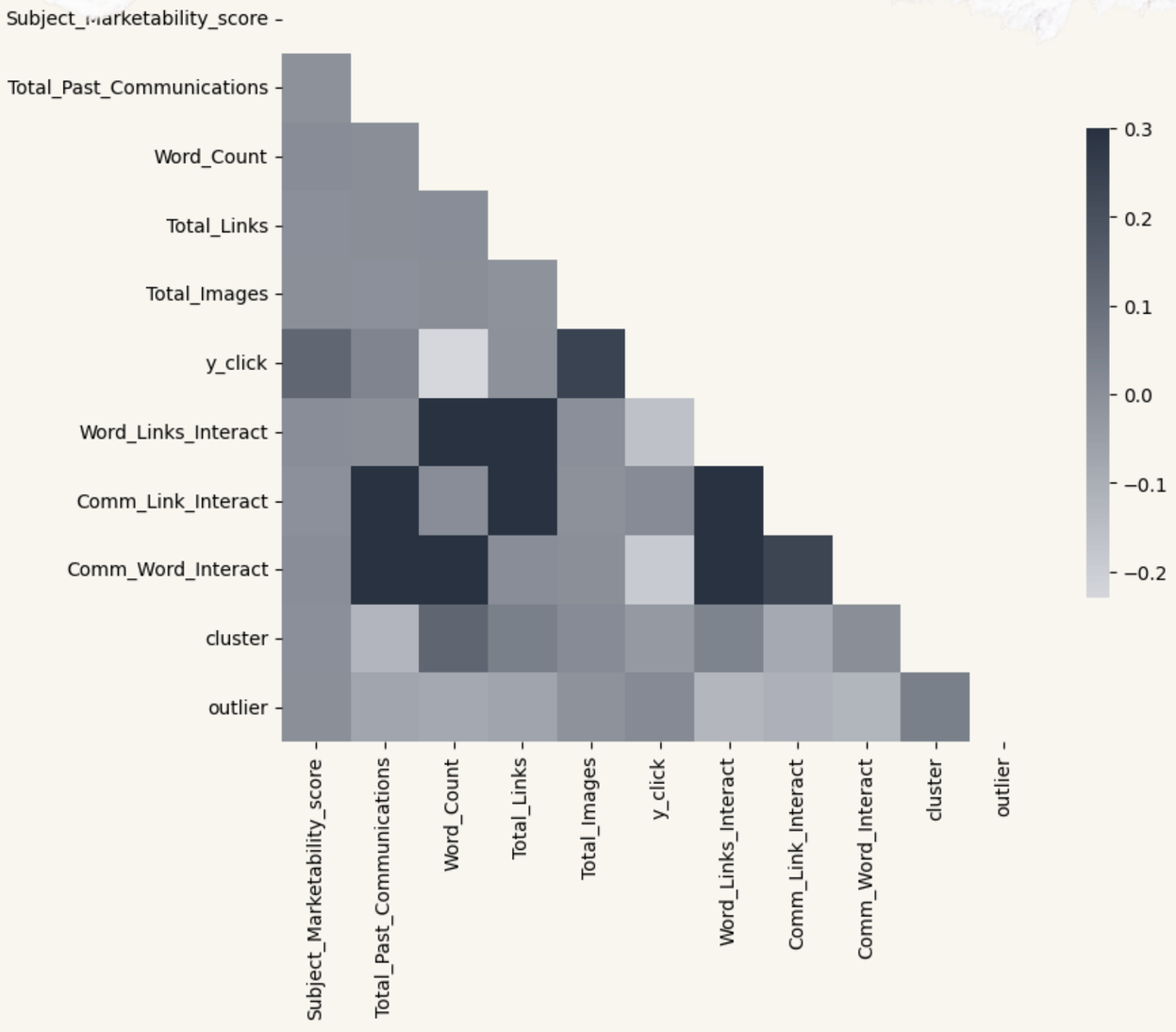
# Cluster Characteristics




By examining the mean values, we identify distinguishing features and commonalities:

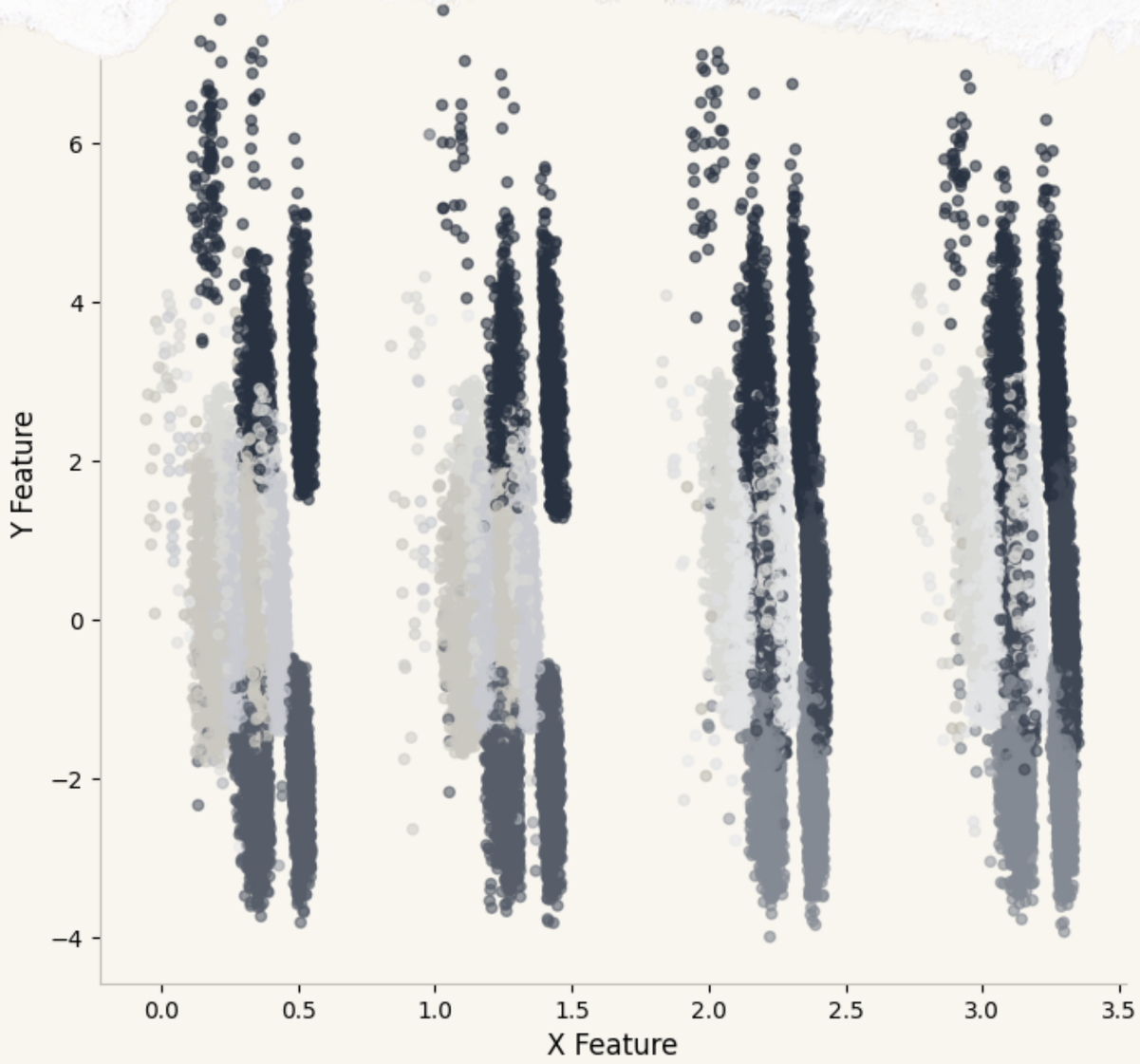
- 'Total_Past_Communications' value fluctuates between 9.79 and 15.61, suggesting different communication histories across clusters.
- The 'Word_Count' feature varies significantly, with certain clusters favoring concise or verbose content.

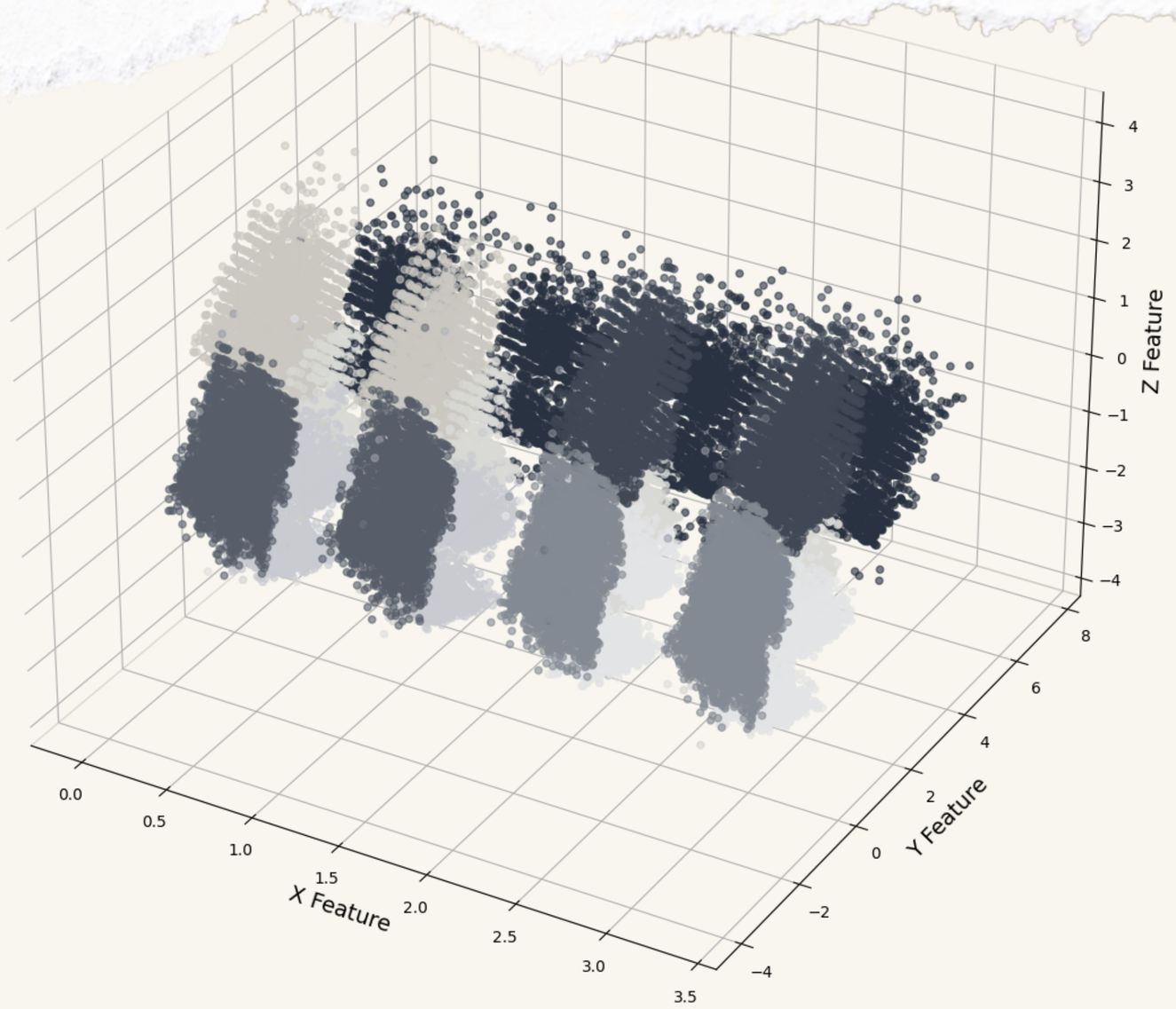
# Feature Correlations Explored



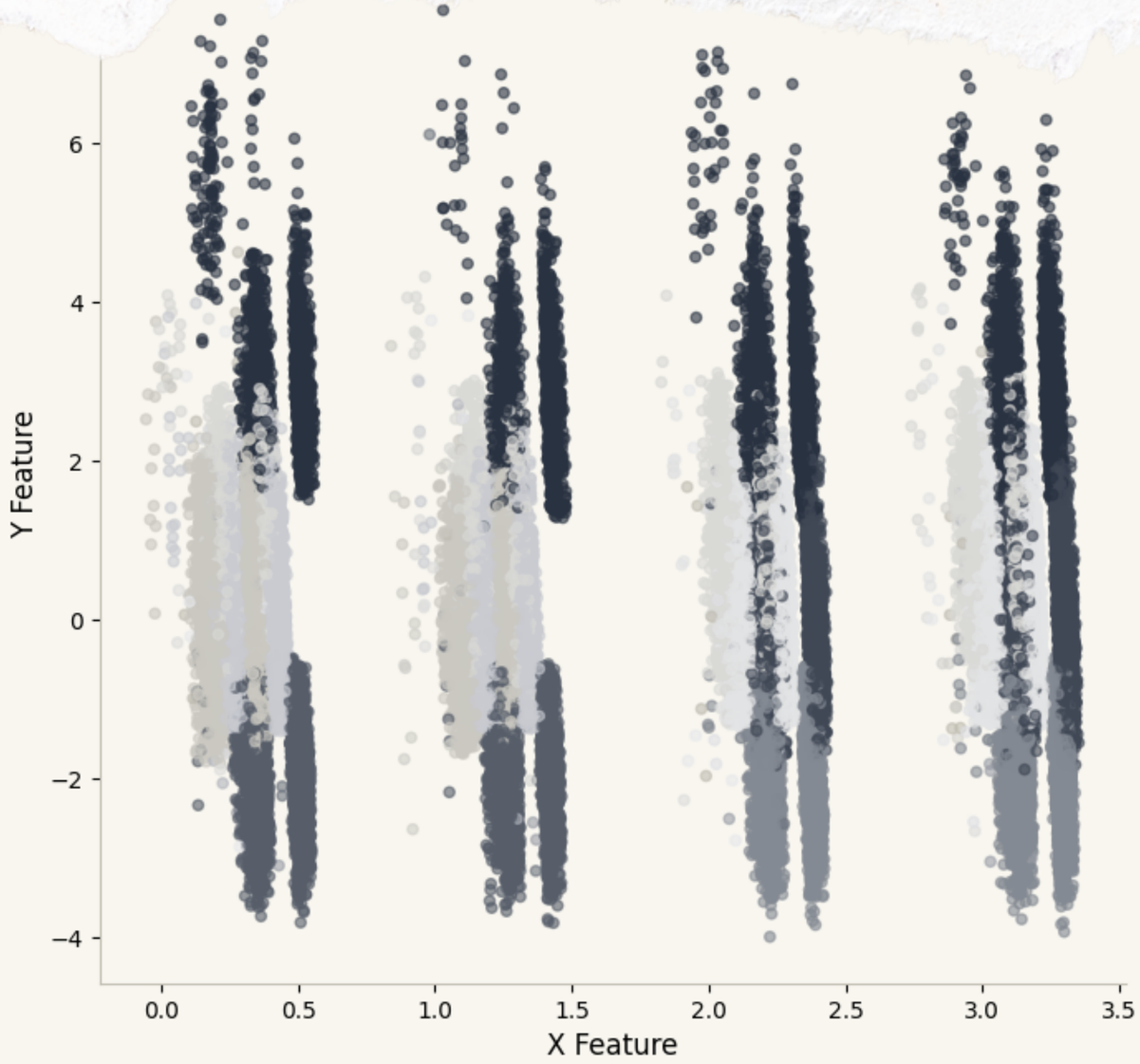

- There is a positive correlation between the number of clicks ('y_click') and 'Total_Images', suggesting that visual content might boost engagement.
- Conversely, a high 'Word_Count' correlates negatively with 'y_click', indicating a propensity to skip text-heavy content.
- These insights shed light on key relationships within our data and guide further feature engineering and content strategies.

# Truncated SVD Representation



# Truncated SVD Representation




SVD enables a **2D representation of our dataset**, revealing distinct strip-like formations.

Each strip comprises four clusters, with **color coding i**ndicating the eight major clusters.

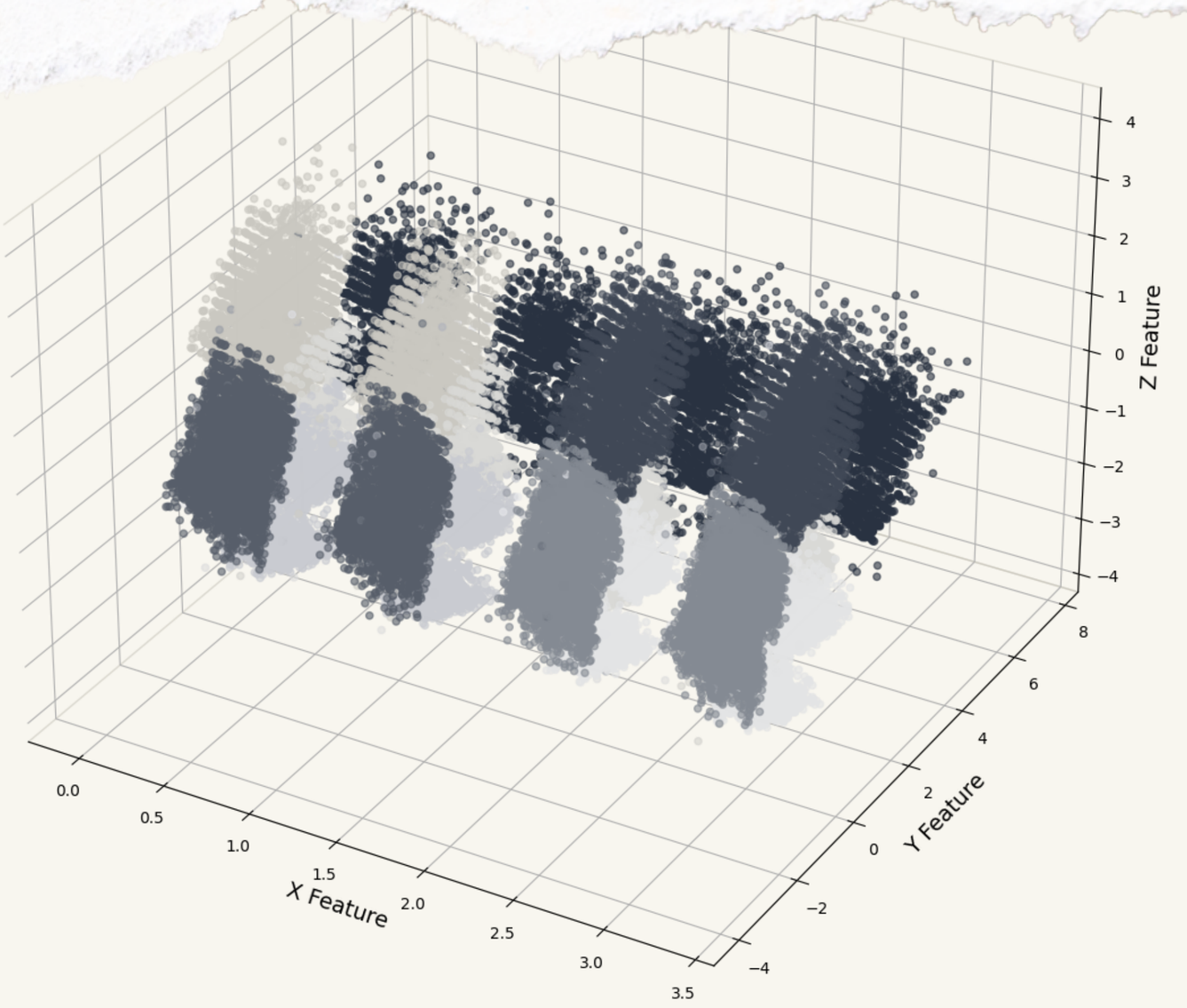
This visualization offers a unique perspective on our data's structure and the distribution of our clusters.

# Truncated SVD Representation

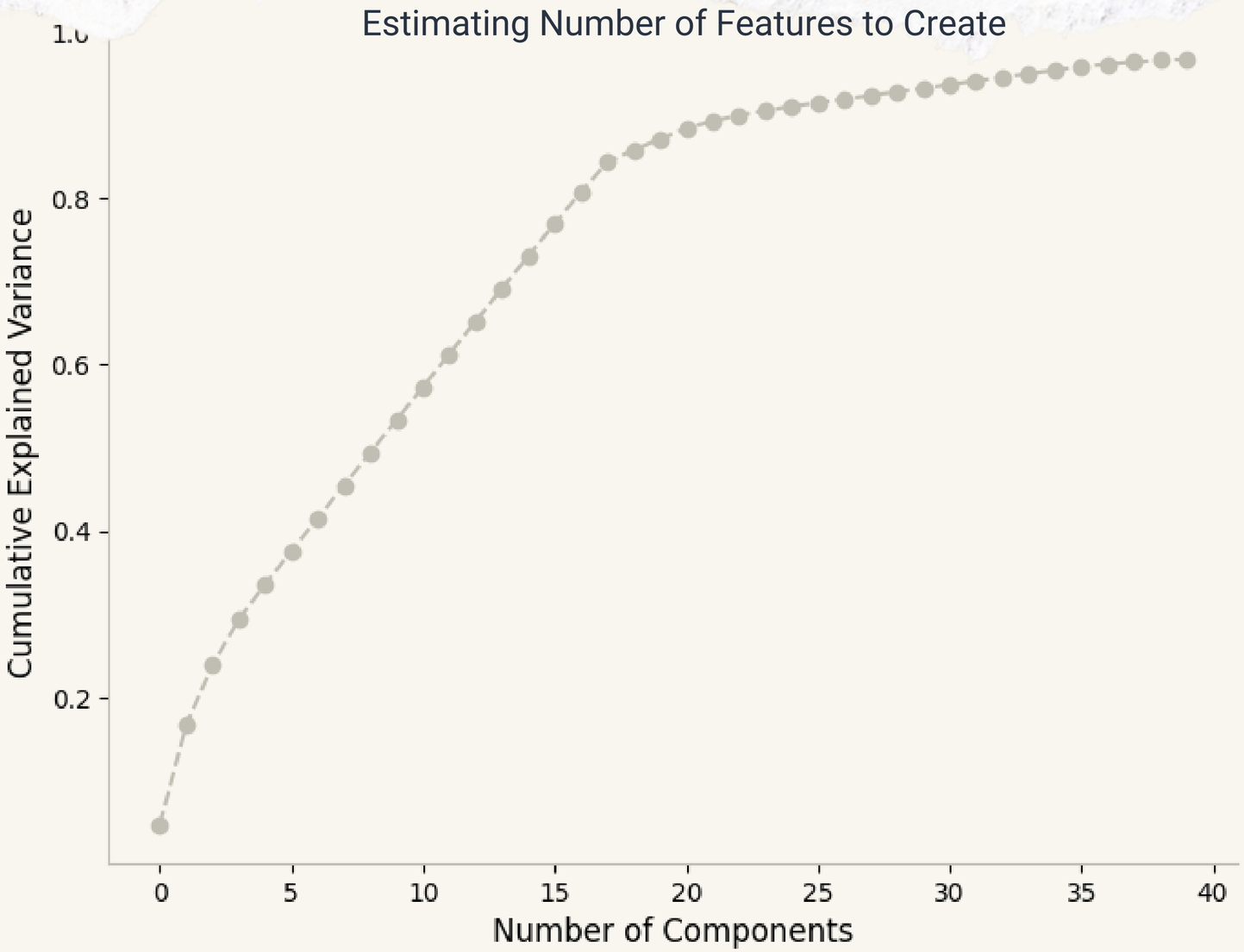
The **3D SVD representation** of our dataset reveals a formation resembling four ovals, each divided into sectors representing clusters.

Although the specific features produced by SVD can't be explicitly interpreted, the structured plot confirms our **data's meaningful relationships**.

The visual distinction between clusters reinforces the representativeness of this feature.



# Creating Features with SVD




Despite a common tendency for dimensionality reduction methods to decrease metric results, increasing the number of features in this case improved model performance.

By calculating cumulative variance, we determined the number of SVD components to explain the majority of the data, choosing **20 components** that explained approximately **87% of the dataset.**

# Models Grid Search

Train/validation/test splits on validation data with additional cross-validation, using recall as scoring. The split is 80/10/10.

Gradient Boosting models, especially LightGBM, performed better than any other model in **precision**, **recall**, **accuracy**, and **speed**.

Despite the tendency to overfit, highly-unbalanced data requires more complicated models.

96

Models were fitted

6

Classifiers types were tested

- Logistic Regression
- SGD Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost
- LightGBM

4

Steps pipeline introduced

1. Preprocessing (Encoding and Scaling)
2. Oversampling (KMeansSMOTE)
3. Modeling
4. Cross-Validation on unseen data

# LightGBM Best Model

98.2%

Cross-Validation (5 folds)
average **recall** in Grid Search

96.7%

Cross-Validation (5 folds)
average **recall** on Validation Set

93.3%

**Accuracy** on Test

50.3%

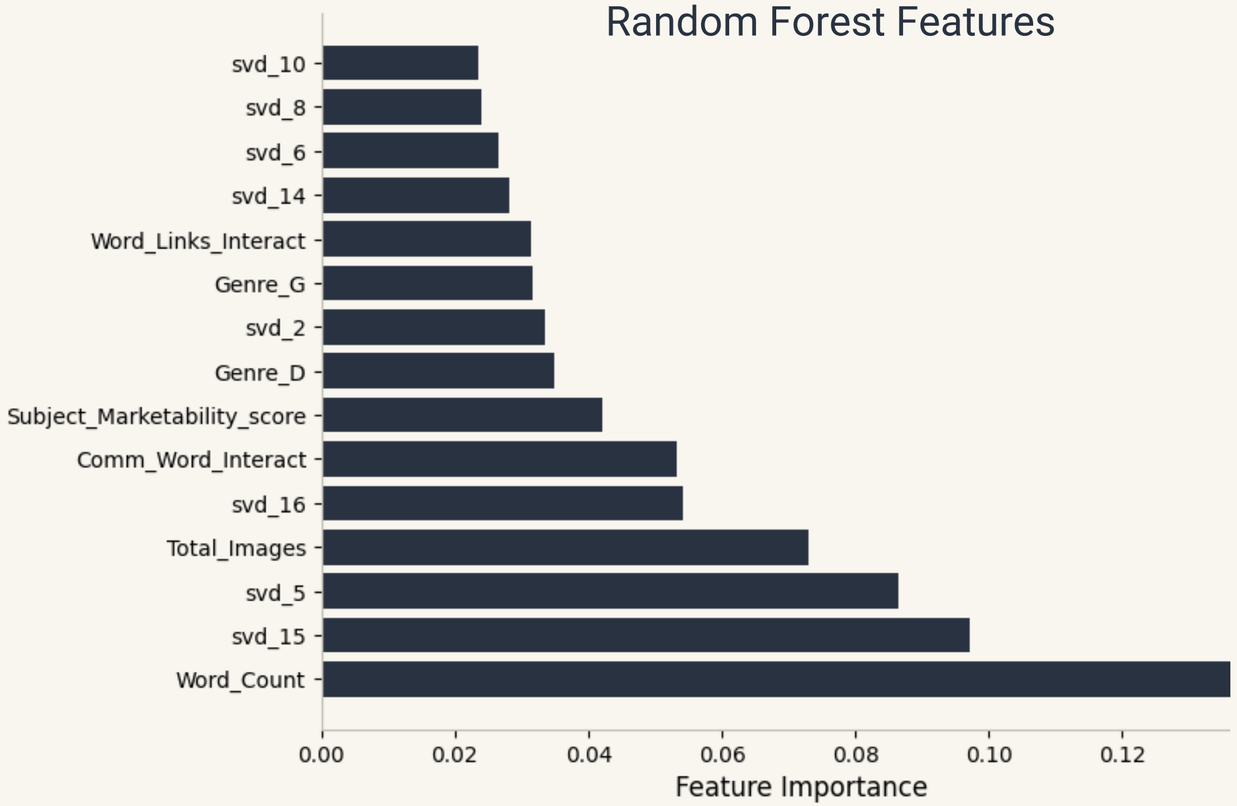
**Recall** on Test

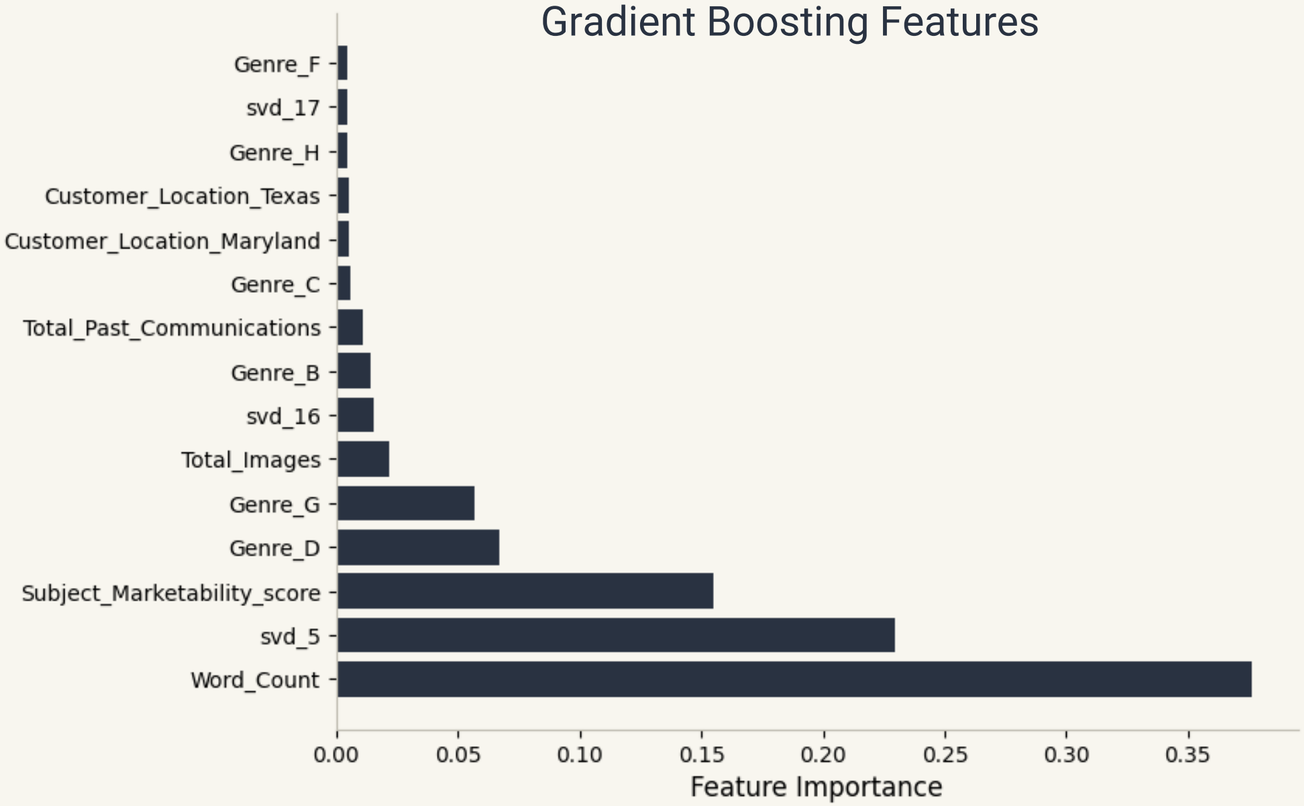
99.1%

**Precision** on Test

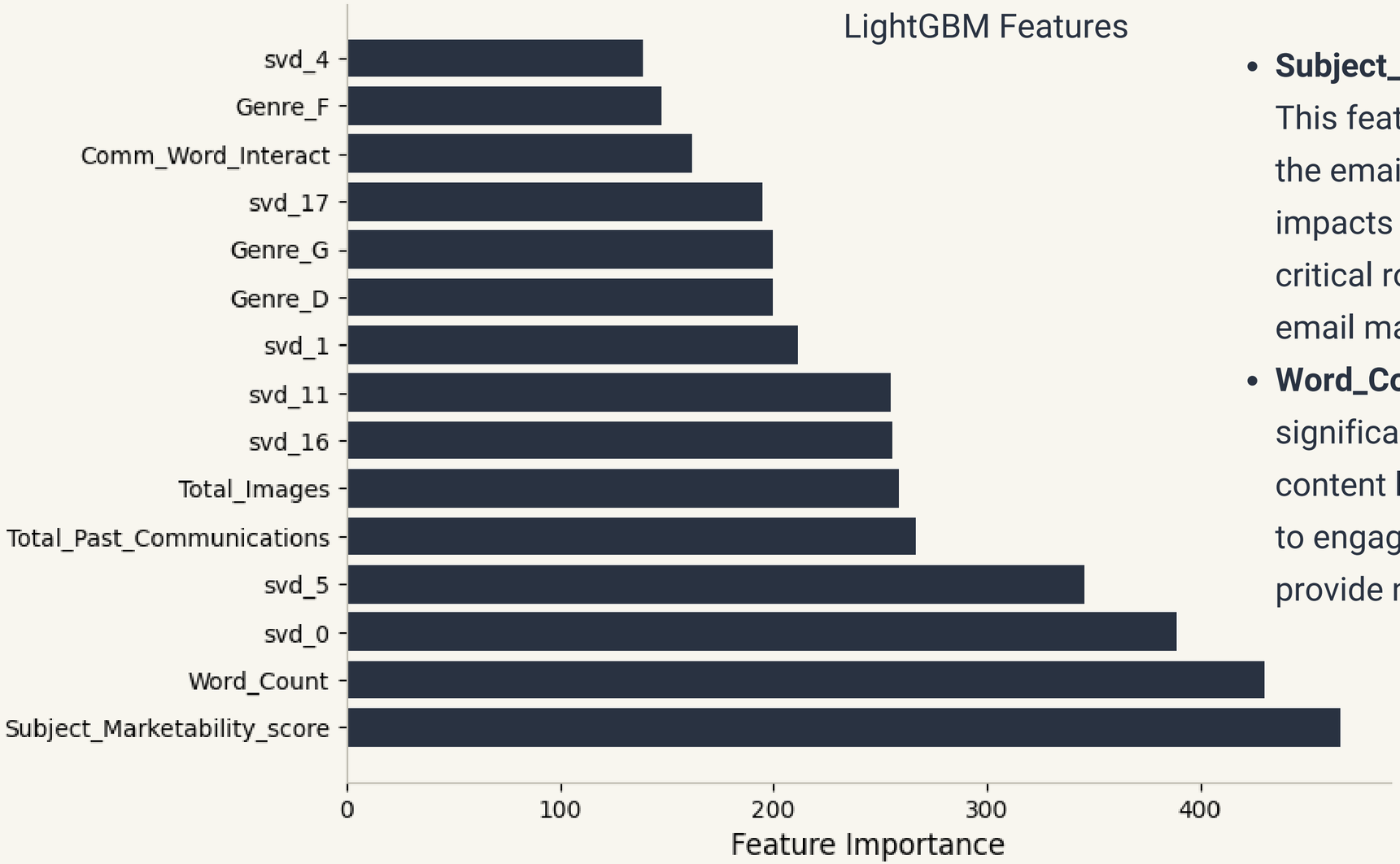
# Features Importance

Examples for other classifiers



# Features Importance




- **Subject_Marketability_score (Importance: 466):** This feature, as the most important, indicates that the email subject's attractiveness significantly impacts whether an email is clicked. This shows the critical role of crafting engaging subject lines for email marketing.
- **Word_Count (Importance: 430):** The second most significant feature indicates the importance of email content length. It suggests that recipients are likelier to engage with emails that balance brevity and provide necessary information.

# Major Insights

**Data Complexity & Feature Engineering:** The data's complexity was successfully tackled by creating 44 new features using clustering, an Isolation Tree for outlier detection, and accounting for missing values in marketing columns.

**Impact of Dimensionality Reduction:** Using SVD for visualization revealed meaningful patterns and demonstrated the importance of feature creation, even though it initially seemed counter-intuitive given the large number of features.

**Model Selection & Performance:** Out of almost a hundred models fitted, LightGBM outperformed others, exhibiting a balance between precision, recall, and computation speed.

**Significance of Specific Features:** The Subject_Marketability_score and Word_Count proved to be the most influential features in predicting email click-throughs.

# Conclusion

- Through comprehensive data analysis, we've gained **valuable insights** into the features influencing email click-throughs, notably the Subject_Marketability_score and Word_Count.
- The **LightGBM model** proved the **most effective** tool for our task, demonstrating the need for complex models when dealing with highly imbalanced data.
- Generating **44 new features** improved the model's performance and provided a richer dataset understanding.
- Despite the challenges, the insights gained offer promising future strategies to improve email engagement.