Analysis and Prediction of Email Click-Through Rate

Chae Uk Lim
The Department of Computer Science
University of North Texas
Denton, Texas, USA
ChaeUkLim@my.unt.edu

Lakshmi Keerthi Gopireddy
The Department of Computer Science
University of North Texas
Denton, Texas, USA
LakshmiKeerthiGopireddy@my.unt.edu

Kiran Jyothi Bodduluri
The Department of Computer Science
University of North Texas
Denton, Texas, USA
KiranJyothiBodduluri@my.unt.edu

Sajid Ali Shaik
The Department of Computer Science
University of North Texas
Denton, Texas, USA
SajidAliShaik@my.unt.edu

Siri Chandana Siripurapu
The Department of Computer Science
University of North Texas
Denton, Texas, USA
SiriChandanaSiripurapu@my.unt.edu

***Abstract*—Today, we have multiple enterprises and numerous products in every field, resulting in a highly competitive environment where promoting business and reaching relevant customers is a significant task for all companies. With the increase in the use of digital equipment and online content, displaying advertisements digitally to customers has become a common form of product promotion. But, we can neither promote the same product to the entire world population nor all the products to a single customer. So, we need to understand whether an advertisement is relevant to a customer and analyze or predict the probability of the customer consuming the advertisement. This motivated us to analyze the relationship between product-related attributes and customer data, alongside predicting the probability of a customer clicking the advertisement.**

I. Introduction

The anticipation of click-through rates stands as a pivotal tool for enterprises, enabling precise customer targeting and amplification of sales. This project primarily focuses on analyzing the click-through rates affiliated with advertisements showcased within email platforms.

The primary objective of this study is to scrutinize the variables influencing customer engagement with specific advertisements and substantiate these influences through empirical statistical analysis. Leveraging these identified variables impacting the click-through rate, the project endeavors to develop a machine learning model aimed at forecasting the likelihood of customers clicking on individual advertisements.

The study encompasses several key elements:

- Exploratory Data Analysis
- Correlation between features and target
- Visualization to understand the relation between each feature and target
- Statistical test to prove the relation between each feature and target
- Data distribution of target variable using Bootstrap sampling method
- Training a model using selected features
- Perform principle component analysis
- Re-train and analyze performance
- Train model using Cross Validation

II. Related Work

[1] This paper discusses about click-through rate predictions of Google advertisements available on the YouTube website or application. This paper uses a deep learning model to predict the click-through rate. [2] This paper discusses about the prediction of clicks on advertisements on Facebook.

III. Dataset

The utilized dataset originates from Kaggle and pertains to the click-through rates observed within email-delivered advertisements.[3] This comprehensive dataset comprises 21 columns delineating various attributes related to advertisement details. The distinctive features of this dataset encompass:

- 'campaign_id': A distinctive identifier assigned to each displayed advertisement.
- 'sender': Entities or enterprises responsible for disseminating email advertisements.
- 'category': Classification or thematic categorization of content showcased within the email.
- 'product': Specific products promoted or mentioned in the advertisement.
- 'day_of_week': The specific day when the email containing the advertisement is received.
- 'is_weekend': A binary indicator signifying receipt of the email on a weekend.
- 'times_of_day': Segmentation based on the temporal context of email reception (e.g., Morning, Evening, Night).
- 'no_of_CTA': Count of Call-to-Action (CTA) elements embedded within the email content.
- 'mean_CTA_len': Average length of the Call-to-Action elements within the email.

- 'is_image': Indication of the presence of images within the email.
- 'is_personalised': Binary indicator determining whether the email content is personalized for a specific recipient group.
- 'is_quote': Count of quotations present in the email.
- 'is_timer': Indication of temporal elements included within the email.
- 'is_emoticons': Count of emoticons incorporated in the email content.
- 'is_discount': Indication of promotional discounts featured in the advertisement.
- 'is_price': Displayed price of items within the advertisement, if applicable.
- 'is_urgency': Highlighting whether the email is marked as urgent.
- 'target_audience': Specified intended audience for the advertisement.
- 'subject_len': Length of the subject line in the email.
- 'body_len': Extent of content displayed within the email.
- 'mean_paragraph_len': Average length of paragraphs within the email.

The paramount target variable under investigation is:

- 'click_rate': The Click Through Rate (CTR), derived as the ratio of advertisement clicks to the total count of emails dispatched, serving as a fundamental metric for gauging advertisement efficacy within this dataset.

IV. Details of Feature

In the initial phase of Exploratory Data Analysis (EDA), a uni-variate analysis is conducted to comprehend the dataset's essence. This involves scrutinizing data distributions and summarizing numerical feature statistics through visual aids such as bar plots, scatter plots, or histograms. Concurrently, categorical feature analyses are performed, elucidating feature significance and distribution using bar plots. Identification of outliers and missing values is pivotal, providing insights into the dataset's general characteristics encompassing mean, minimum, and maximum values.

Subsequently, the correlation matrix is constructed to ascertain the relationship between individual features and the target variable. The heatmap visualization method is employed to depict this correlation matrix, presenting feature-to-feature relationships alongside feature-to-target correlations.

Visualizations, including scatter plots and bar plots, are instrumental in comprehending the association between each feature and the target variable. These plots facilitate the swift identification of potential linear relationships within the dataset.

Employing the Bootstrap sampling method, estimations are made concerning the target variable's data distribution. Specifically, the mean click-through rate is calculated, providing a 95% confidence interval for the range within which the mean value falls.

Statistical testing techniques, encompassing correlation coefficients, t-tests, and Ordinary Least Squares (OLS) linear regression methods, are employed to establish the impact of each feature on the click-through rate. These tests ascertain the statistical significance of features in influencing the target variable, aiding in the shortlisting of impactful features.

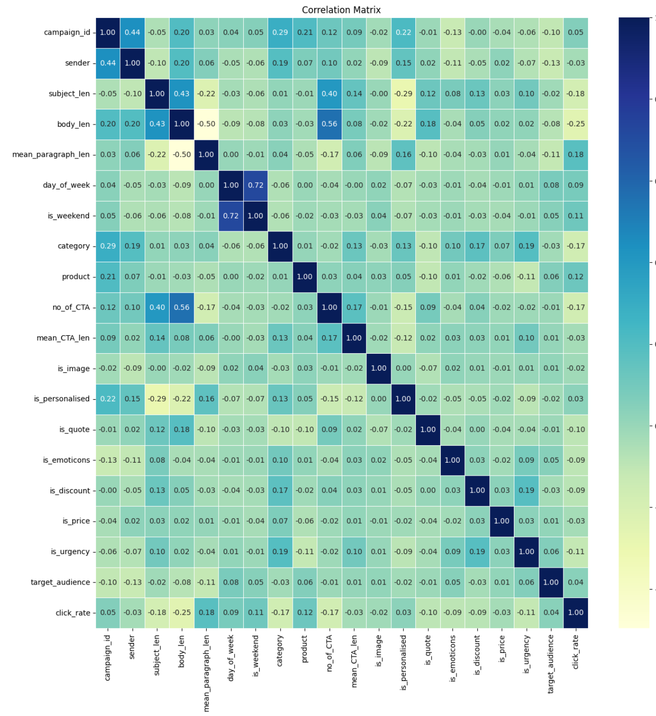


Fig. 1. Correlation Matrix

Post-feature selection via statistical evidence, a machine learning model is constructed utilizing the chosen features. Prior to model training, essential data preprocessing steps are executed, including outlier removal and categorical variable encoding. The model's outcomes are subsequently analyzed to gauge its efficacy in predicting the click-through rate based on the selected features.

V. Analysis

A. Correlation between features and target

To understand the correlation between features and, feature and target, displaying a correlation matrix. From the matrix, we can infer that there is strong relation between the below mentioned pairs: No. of CTA and Body Length No. of CTA and Subject Length Is weekend and Day of week Body Length and Subject Length Campaign ID and Sender

B. Exploratory Data Analysis and Statistical Analysis:

Null hypothesis: There is no relationship between the feature mentioned below and the click-through rate.

Alternate hypothesis: There is a relationship between the feature mentioned below and the click-through rate.

1) campaign id: We observe that the values in this feature are continuous numbers that represent the ID for each column number. This serves as a serial number, so excluding this feature from the analysis.

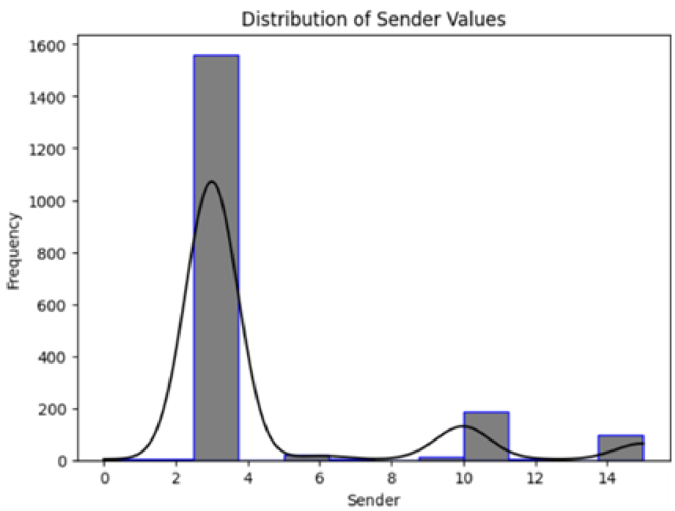


Fig. 2. Distribution of Sender Values

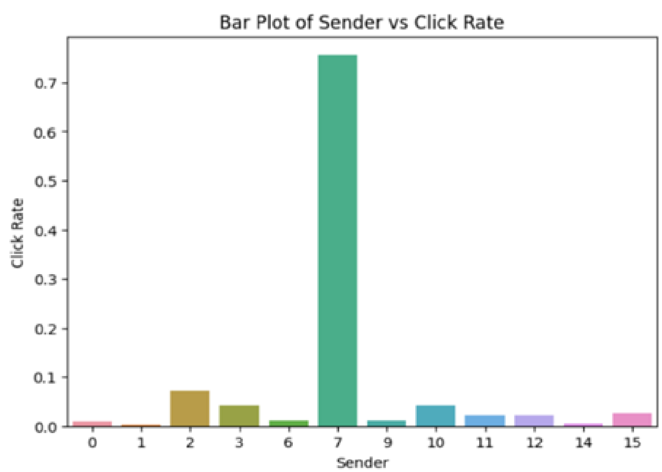


Fig. 3. Box Plot of Sender vs Click Rate

2) sender: The sender refers to the person who sends the email. And there are a total of 1888 values in the range of 0 to 15 and the mean is 4.4 with a standard deviation of 3.28. This plot indicates a histogram, where the x-axis is the sender and the y-axis is the frequency of the sender. From the histogram, we can infer that the range of values is more towards the upper bound and lesser towards the lower bound. To understand the relationship between sender and click-through rate. we are plotting the bar plot with the sender on the x-axis and the click-through rate on the y-axis. We can see that all the senders almost have the same click rate from 0.0 to 0.1 as sender 7 has the highest click rate with 0.8. where most of the senders are of category 3. This indicates that there is a relationship between the sender and the click-through rate.

We can observe from the data visualizations that the subject length affects the click-through rate. We performed t-test and ANOVA tests to find the correlation. The correlation coefficient between the sender and click_rate is 0.0013 as this is near zero, we can say that there is a weak correlation and

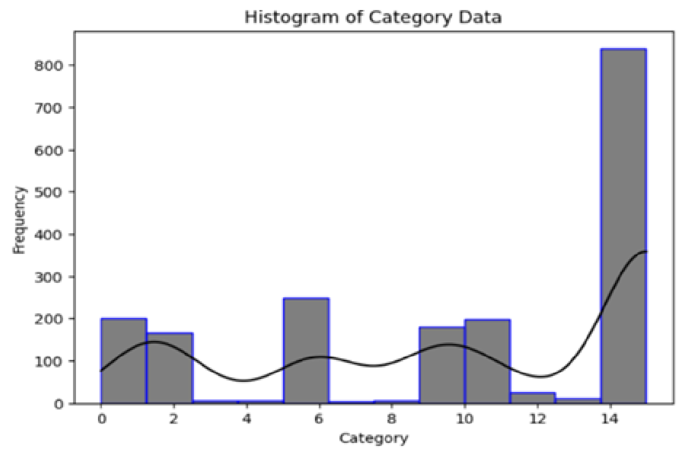


Fig. 4. Histogram of Category

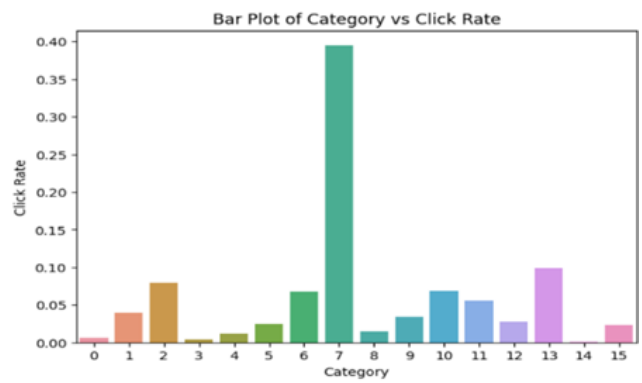


Fig. 5. Bar Plot of Category vs Click Rate

the p-value is 0 according to the t-test. The p-value is less than 0.05 significant level so we can reject the null hypothesis.

3) category: There are 1888 values in total. This is related to the above product column. These categories are of 15 types. The mean is 9.95 and the standard deviation 5.3. This plot indicates a histogram, where the x-axis is the category and the y-axis is the frequency of the category. From the histogram, we can infer that the range of values is more towards the upper bound and lesser towards the lower bound.

To understand the relationship between sender and click-through rate. we are plotting a bar plot with category in the x-axis and click-through rate in the y-axis. The bar plot shows that all the category almost has the same click rate from 0.0 to 0.1 as category 7 has the highest click rate in the range of 0.35 to 0.4. This indicates that there is a relationship between category and click-through rate.

We can observe from the data visualizations that the subject length affects the click-through rate. We performed t-test and ANOVA tests to find the correlation. The correlation between the category and click_rate is 0.0723 as this is near zero, we can say that there is a weak correlation and the p-value is 0 according to the OLS regression model, t-test, and ANOVA. The p-value is less than 0.05 significant level so we can reject the null hypothesis.

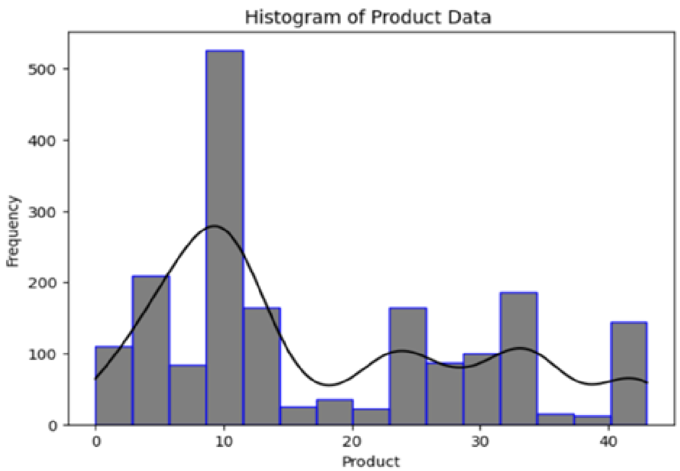


Fig. 6. Histogram of Product

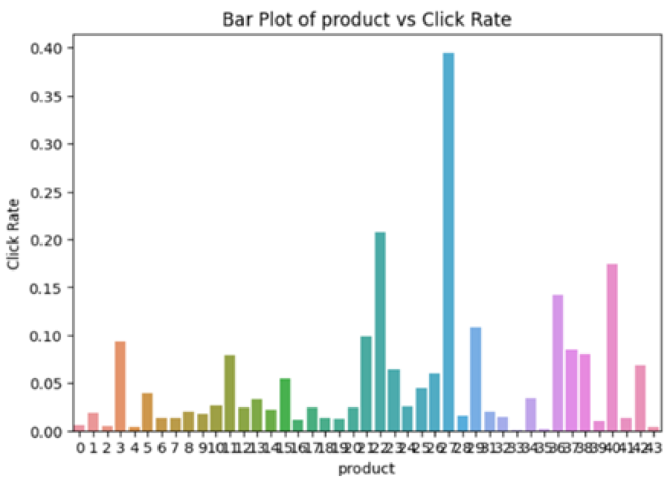


Fig. 7. Bar Plot of Product vs Click Rate

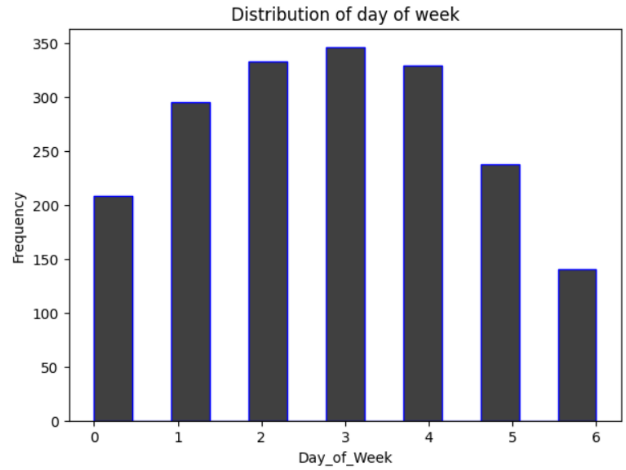


Fig. 8. Histogram of day of week

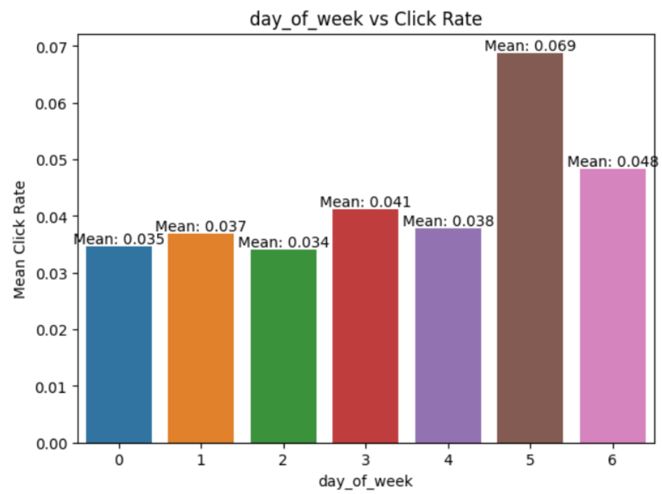


Fig. 9. day of week vs click rate

4) product: There are a total of 1888 values in the range of 0 to 44 the mean is 17.5 and the standard deviation is high at 12.36. this shows that variation is present in the values of the product. This plot indicates a histogram, where the x-axis is the product and the y-axis is the frequency of the product. From the histogram, we can infer that the range of values is more towards the upper bound and lesser towards the lower bound.

To understand the relationship between product and click-through rate. we are plotting a bar plot with the product on the x-axis and click-through rate in the y-axis. The bar plot shows that all the products almost have the same click rate from 0.0 to 0.1 as product 7 has the highest click rate with 0.8. where most of the products are of category 3. This indicates that there is a relationship between product and click-through rate.

We can observe from the data visualizations that the subject length affects the click-through rate. We performed t-test and ANOVA tests to find the correlation. The correlation between the sender and click_rate is 0.0013 as this is near zero, we can say that there is a weak correlation and the p-value is 0 according to the t-test. The p-value is less than 0.05 significant level so we can reject the null hypothesis.

5) day_of_week: The 'days of the week' column is represented numerically from Sunday to Saturday. This numerical representation simplifies computational processes and analysis by providing a standardized numerical scale for the days of the week, ranging from 0 for Monday to 6 for Sunday. It appears from the depicted graph that the volume of emails dispatched declines notably during the weekend, reaching its lowest point. Conversely, there is a discernible uptick in email volume as the week progresses towards its midpoint. This trend indicates a rise in email activity during the middle of the week compared to the weekends, reflecting a notable fluctuation in email transmission across the week's duration.

According to the graph above, the click conversion rate of emails sent on Saturday is significantly higher than that of emails sent on other days of the week.

The calculated p-value denoting the relationship between

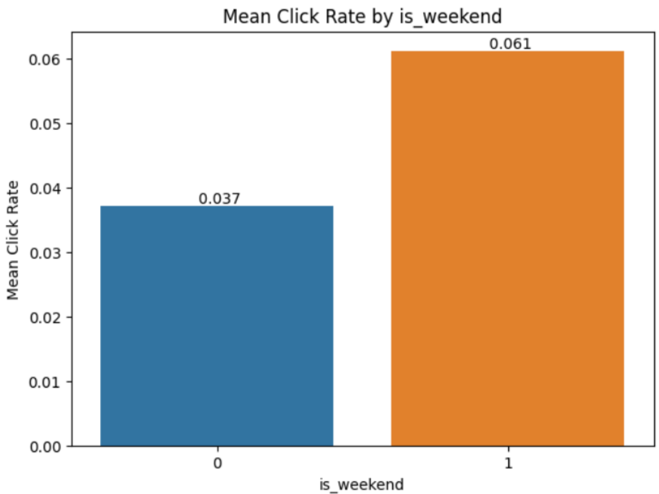


Fig. 10. Mean click rate by is weekend

the given values is notably minute, approaching a level of insignificance that might appear as 0 for practical purposes, although strictly speaking, it is not absolute zero. This statistical outcome strongly advocates against the null hypothesis. Consequently, in accordance with statistical convention, the null hypothesis is aptly rejected in light of this compelling evidence.

6) is_weekend: The 'is_weekend' column, operating as a binary indicator, primarily comprises values categorized as false, encompassing approximately 80% of the dataset. Remarkably, discernible variations in the mean click rates between weekends and weekdays are apparent. Specifically, the mean click rate during weekends notably surpasses that observed during weekdays.

The statistical test, yielding a p-value of 5.413031606085736e-61, underscores the immense significance of differentiating the independent outcome from the dependent one. This minuscule p-value strongly suggests a robust and substantial statistical significance in the disparity between the click rates observed during weekends compared to those witnessed on weekdays.

7) times_of_day: The 'times of day' column exhibits a trinary categorization, comprising three distinct elements: morning, noon, and evening. Notably, akin to the 'days_of_week' column, the statistical analysis for this column yields a p-value nearly approaching 0. This inference signifies a robust and compelling relationship between the time of day and the 'click rate' value. The minuscule p-value indicates a profound statistical significance, suggesting a substantial association between the designated time segments and the observed 'click rate' values.

8) subject_len: There are a total of 1888 data points i.e., we do not have any missing values in this feature. The value ranges from 9 to 265, with a mean of 86.25. This plot indicates a histogram, where the x-axis is the length of the subject and the y-axis is the frequency of the length. From the histogram,

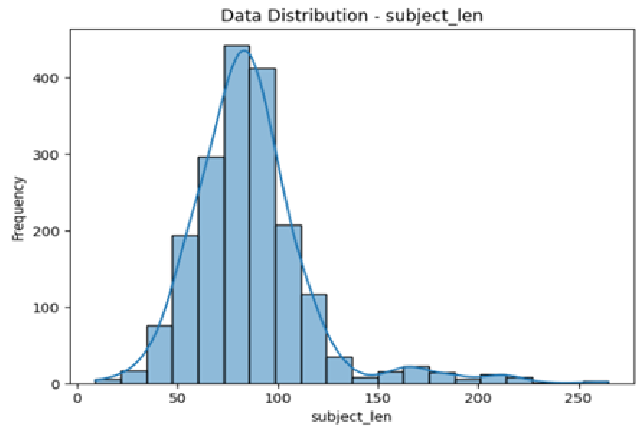


Fig. 11. Data distribution of subject length

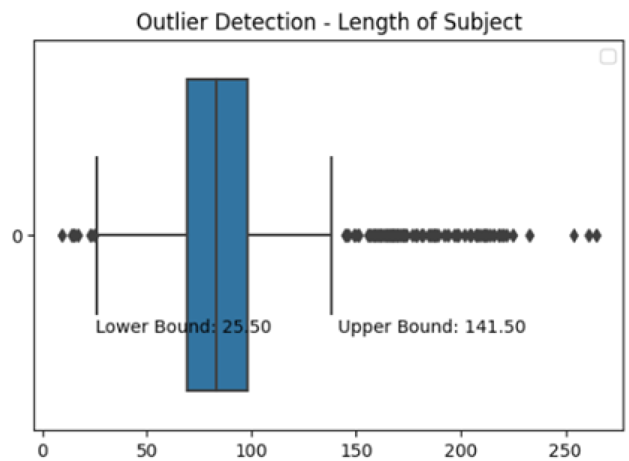


Fig. 12. Subject Length

we can infer that the range of values is more towards the upper bound and lesser towards the lower bound. The Box plot indicates the distribution of data and the outliers. We can observe that there are multiple outliers above the upper bound.

To understand the relationship between length of the subject and click through rate, we are plotting a scatterplot with length of the subject in x-axis and click through rate on the y-axis. We can see the distribution of data is concentrated at a place and there are few points scattered. This indicates that there is relationship between subject length and the target variable.

We can observe from the data visualizations that the subject length effects the click through rate. We performed t-test by using ttest_ind function, and observe that the correlation coefficient is -0.18 and p-value is 3.1 e*-15. The p-value is very low, which indicates that the null hypothesis should be rejected. This gives statistical proof that the subject length might effect the click through rate. Hence, we can consider this feature to predict the click through rate. The correlation coefficient suggest that subject length is negatively correlated to click through rate.

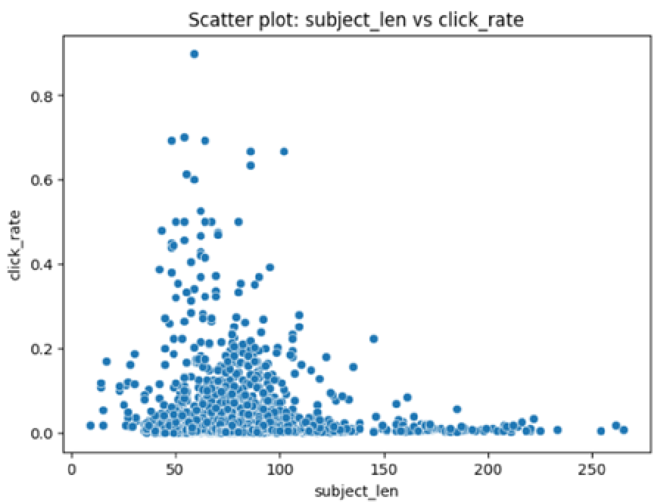


Fig. 13. Scatter Plot: Subject len vs Click rate

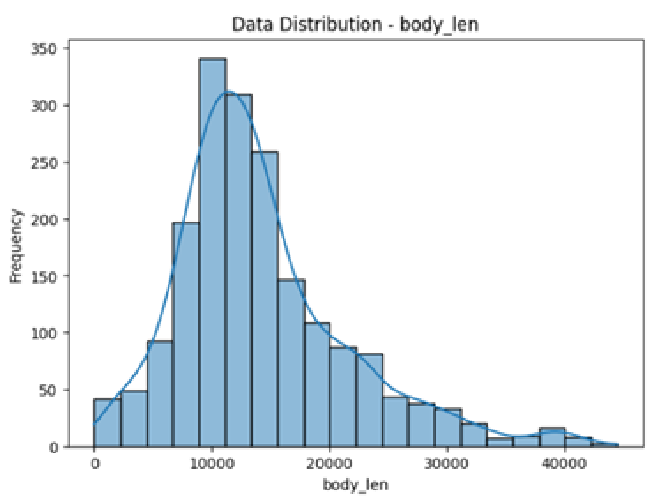


Fig. 14. Data Distribution of body length

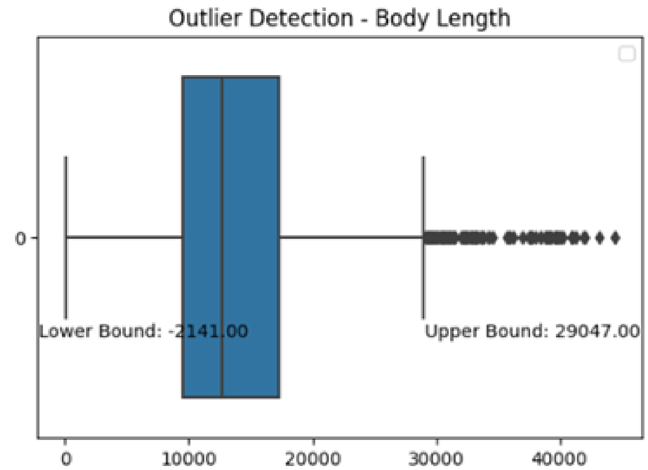


Fig. 15. Outlier detection - body length

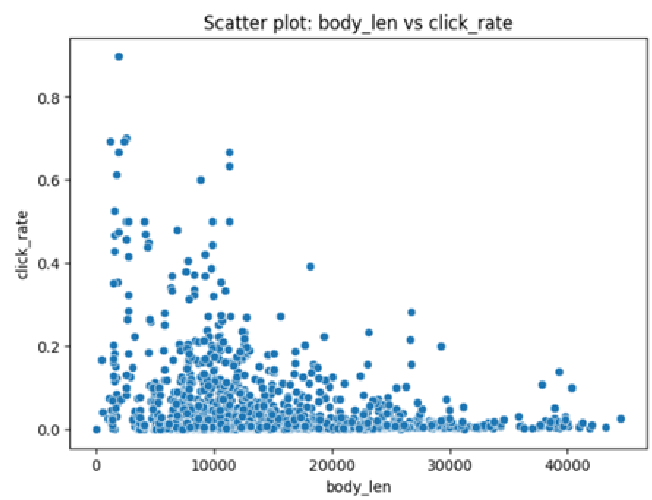


Fig. 16. Scatter plot: body len vs click rate

9) body_len: There are a total of 1888 data points in this feature, which means there are no missing or NA values. The value of this feature ranges from 23 to 44491, with a mean of 14185.78. The below plot is a histogram, where the x-axis is the length of the email body and y-axis is the frequency of the length. From the histogram, we can infer than the range of values are more towards the upper bound and lesser towards the lower bound. The Box plot indicates the distribution of data and the outliers. We can observe that there are multiple outliers above the upper bound.

To understand the relationship between the length of the email body and the click-through rate, we are plotting a scatterplot with a length of the body in the x-axis and click-through rate on the y-axis. We can see the distribution of data is concentrated at a place towards the origin and there are few points scattered. This indicates that there is relationship between body length and click-through rate.

From the above visualizations, we can see that there is visible relation between body length and click through rate. We performed t-test by using ttest_ind function, and observe that the correlation coefficient is -0.247 and p-value is 7.93 e*-28. The p-value is very low, which indicates that the null hypothesis should be rejected. This indicates there is relationship between body length and click through rate. The correlation coefficient suggest that body length is negatively correlated to click through rate.

10) mean_paragraph_len: There are a total of 1888 data points in this feature, which means there are no missing or NA values. The value of this feature ranges from 4 to 286, with a mean of 35.23. The below plot is a histogram, where the x-axis is the mean paragraph length and y-axis is the frequency of the length. From the histogram, we can infer than the range of values are more towards the upper bound and lesser towards the lower bound. The Box plot indicates the distribution of data and the outliers. We can observe that there are multiple outliers above the upper bound.

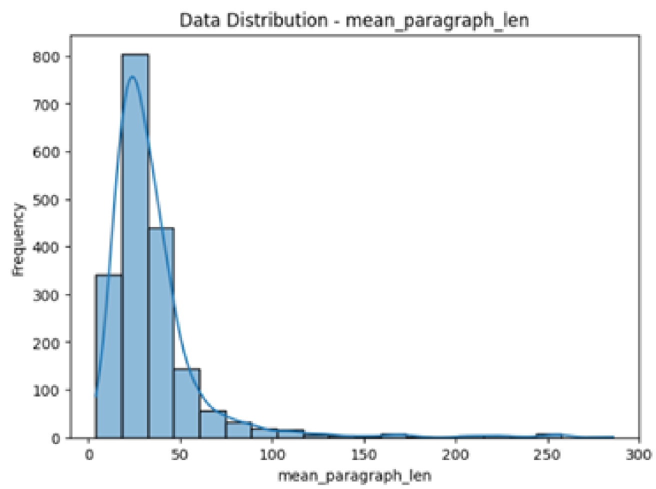


Fig. 17. Data Distribution - mean paragraph len

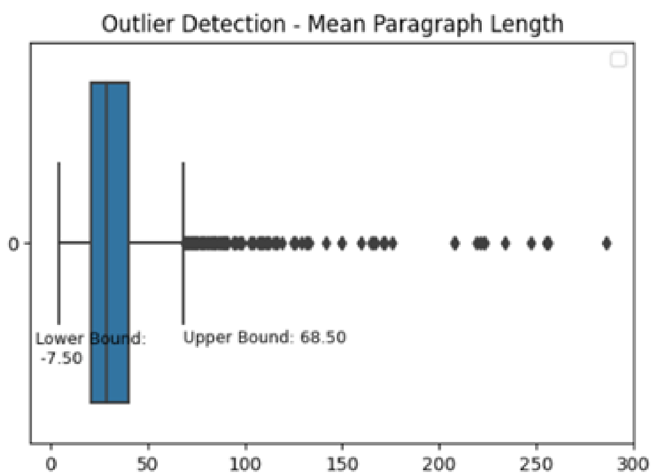


Fig. 18. Outlier Detection - mean paragraph len

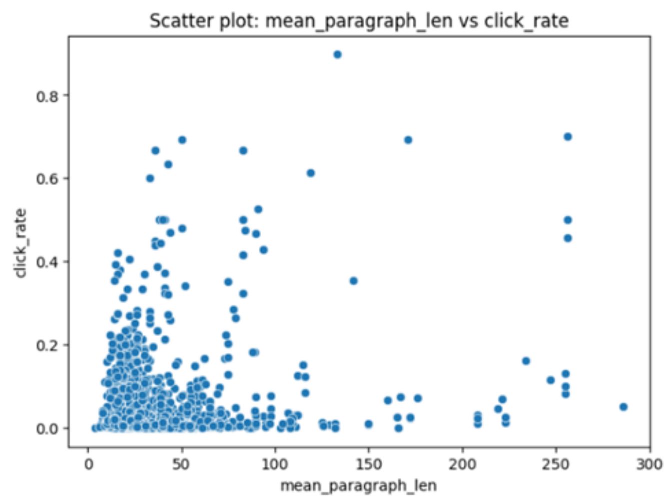


Fig. 19. Scatter Plot: mean paragraph len vs click rate

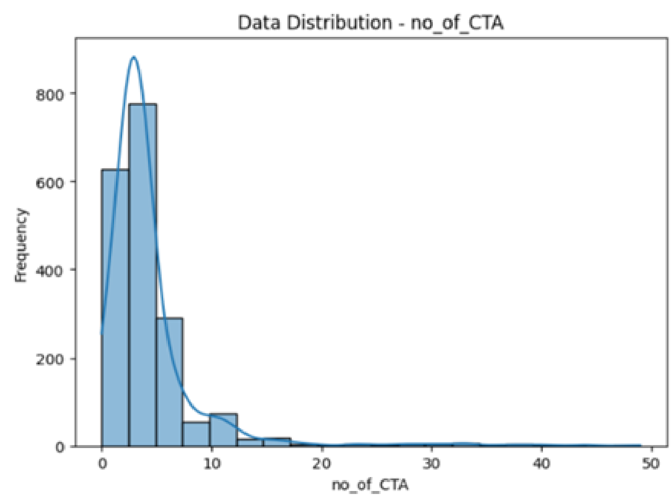


Fig. 20. Data Distribution - no of CTA

To understand the relationship between mean paragraph length and click through rate, we are plotting a scatterplot with mean paragraph length in x-axis and click through rate on the y-axis. We can see the distribution of data is concentrated at a place towards the origin and there are few points scattered. This indicates that there is relationship between body length and click through rate.

From the scatter plot in the visualization, we see that there is relation between mean paragraph length and click through rate. We performed t-test by using ttest_ind function, and observe that the correlation coefficient is 0.17 and p-value is 6.25 e*-15. The p-value is very low, which indicates that the null hypothesis should be rejected. This indicates there is relationship between mean paragraph length and click through rate. The correlation coefficient suggest that mean paragraph length is positively correlated to click through rate.

11) no_of_CTA: There are total of 1888 values and there are no missing values in the data. and this feature values lies in between 0 to 49. The mean of the values is 4.2 and the standard deviation is 4.62. the data distribution of this feature is represented with histogram with no_of_CTA values on x-axis and frequency of the values on y-axis. By seeing the plot the lower bound has more values and there are less towards upper bound. There are some outliers in the data of no_of_CTA above the upper bound. The correlation coefficient is approximately -0.173. This negative value suggests a weak negative correlation between the number of calls-to-action and click rates. In other words, as the number of calls-to-action increases, click rates tend to slightly decrease.

To get a view about the relationship between the number of CTA and click-through rate, we plotted a scatterplot with the number of CTA in the x-axis and click_rate on the y-axis. It is visible that the distribution is mostly concentrated at the origin from 0 to 10 and the remaining points are scattered. There might be a relationship between these features.

As we have observed relation between no_of_CTA variable and click through rate from the visualizations, we are per-

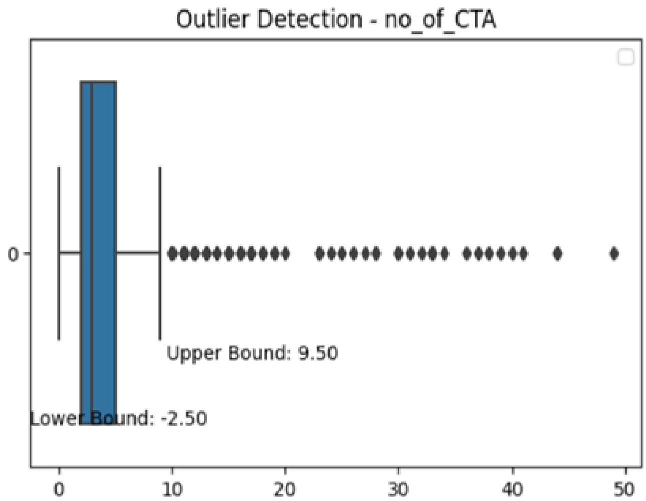
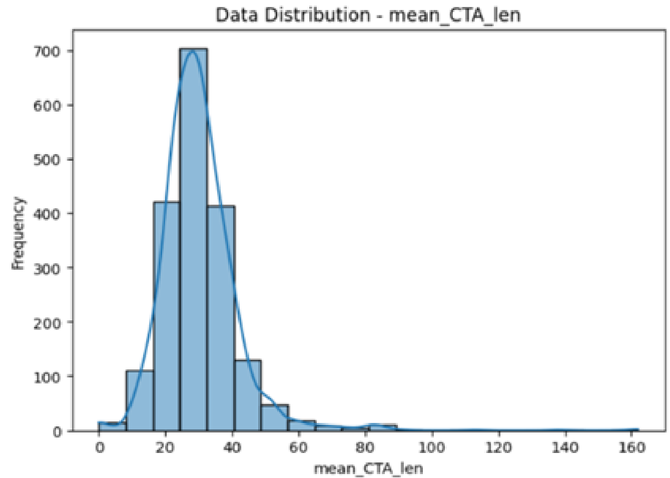
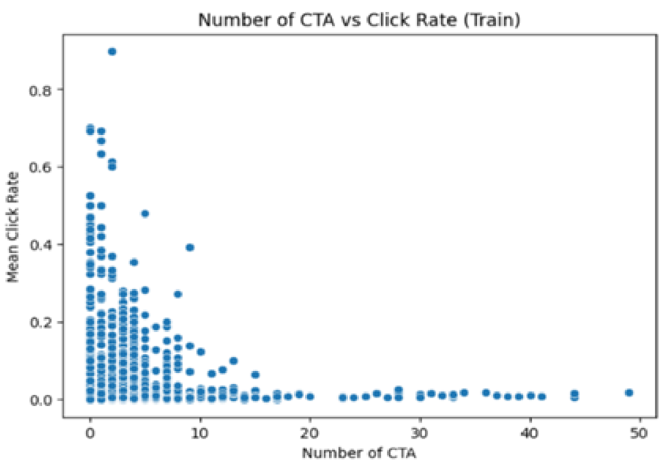
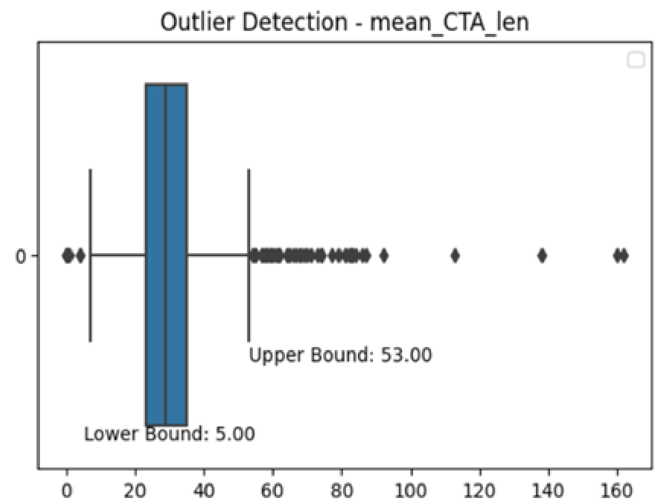


Fig. 21. Outlier detection - no of CTA







forming t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we expect some relation between the feature and click through rate. We can consider this feature to predict the click through rate.

12) mean_CTA_len: There are total of 1888 values in the data and there are no missing values in the data. the mean of values is 30.2 and the standard deviation is 11.8. the data distribution of mean_CTA is represented using histogram. The mean_CTA length is on x-axis and the frequency of the length is on y-axis. We can observe that the range of values are more towards upper bound. And there are some outliers above the upperbound. There are few outliers which are lower than the lower bound.

We have plotted a scatter plot to visualize the relationship between the mean CTA length and click through rate. And this plot has the mean CTA length in x-axis and click rate in y-axis. Here in this plot the data points are mostly at one region between the range 10 to 50 with a average click_rate from 0 to 0.2. And the remaining values are scattred in different regions. There might be a relationship between the mean CTA length and click-through rate.

As we have observed relation between mean_CTA_len variable and click through rate from the visualizations, we are performing t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we expect some relation between the feature and click through rate. We can consider this feature to predict the click through rate.

13) is_image: In this feature there are total of 1888 values which represents that the image is available in the email or not. The mean of the values is 0.91 as the values lies in between 0 to 6. their standard deviation is 0.87. if we observe the barplot we can see that most of the values in the is_image feature are 0's and 1's and there are less images near the maximum value. To visualize the relationship between this feature and click_rate we used a bar plot with is_image on x-axis and click_rate on y-axis. These values lies in between 0 to 6. And most of the values are at category 3 with 0.054 click_rate. There might be a relationship between is_image, click_rate

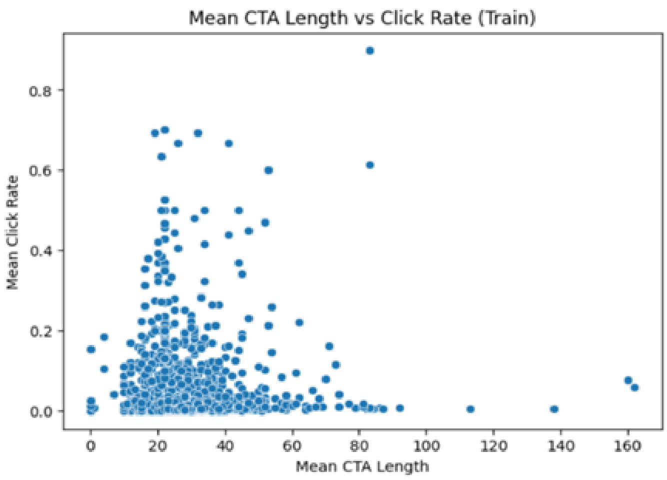
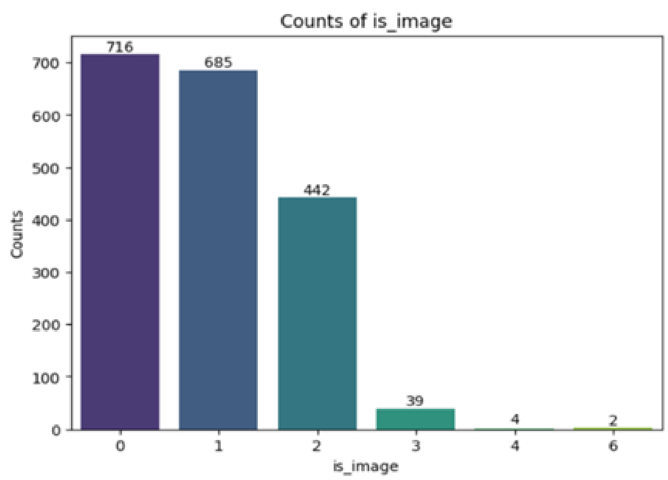


Fig. 25. Mean CTA Length vs Click Rate



and fourth category has the least click through rate of 0.006.

As we have observed relation between is_image variable and click through rate from the visualizations, we are performing t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we expect some relation between the feature and click through rate. We can consider this feature to predict the click through rate.

14) is_personalised: In this feature there are total of 1888 values which represents that the email is personalised or not. The values are 0 and 1. The mean of this feature is 0.06 and the standard deviation is 0.23. Most of the values are of 0's and there are very less 1's. We used pie chart to represent this data as this is a binary data. 94% of the data is having the value as 0 and 6% of the data is 1.

To visualize the relationship between the is_personalised and the click_rate features, we plotted a bar plot with is_personalised on x-axis and click_rate on y-axis. These values are only zeroes and ones with click rate of 0.041 and 0.053 respectively. There is a relationship between the is_personalised fature and click through rate.

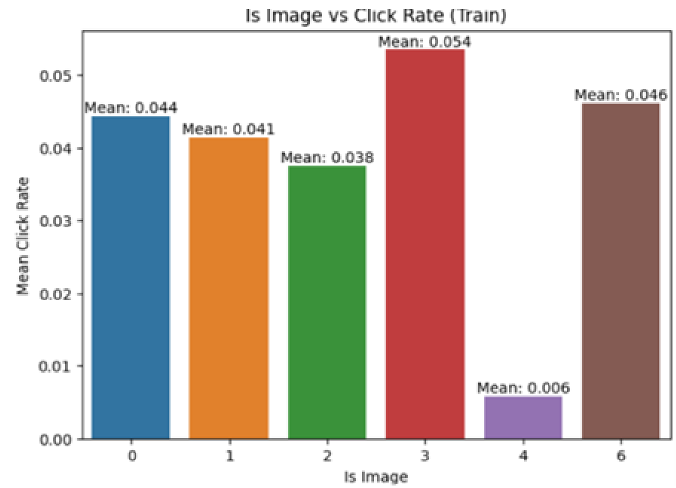
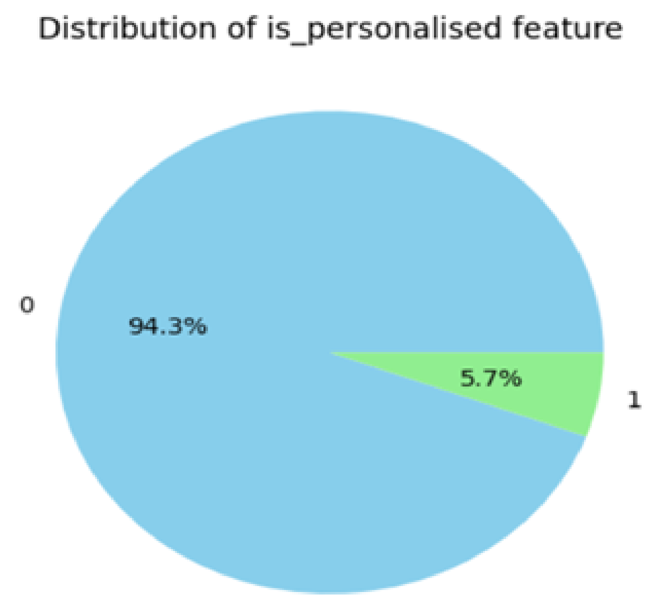


Fig. 27. Is Image vs Click Rate



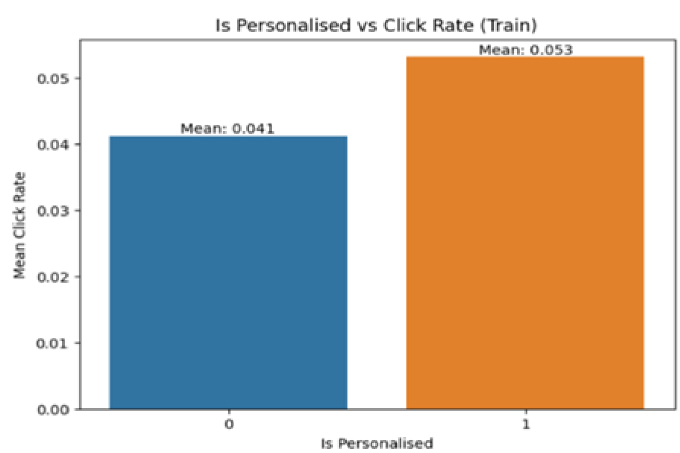


Fig. 29. Is personalised vs Click Rate

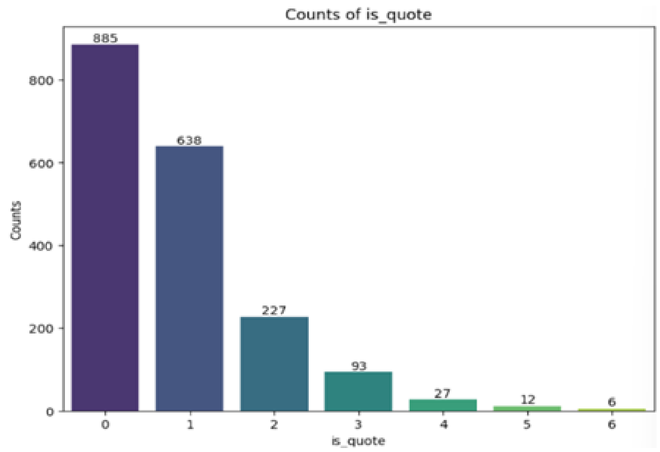


Fig. 30. Count of is quote

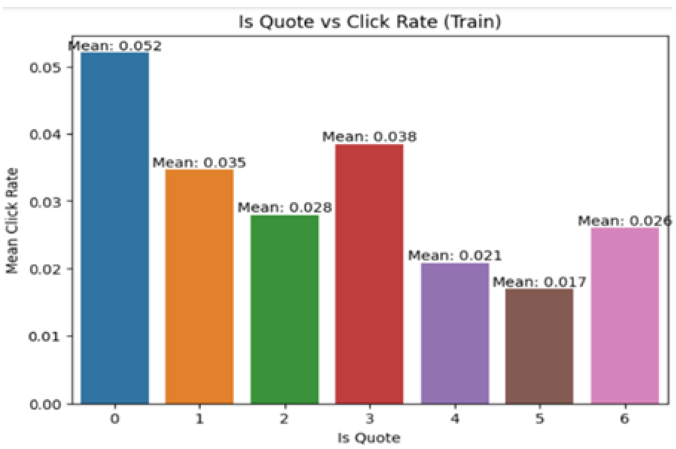


Fig. 31. Is quote vs click rate

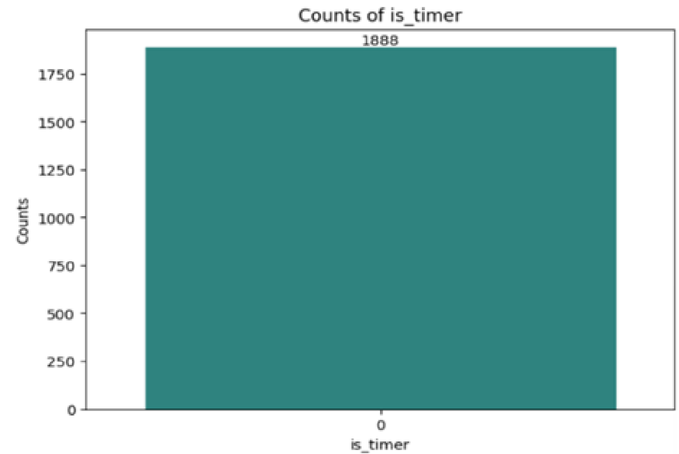


Fig. 32. Count is timer

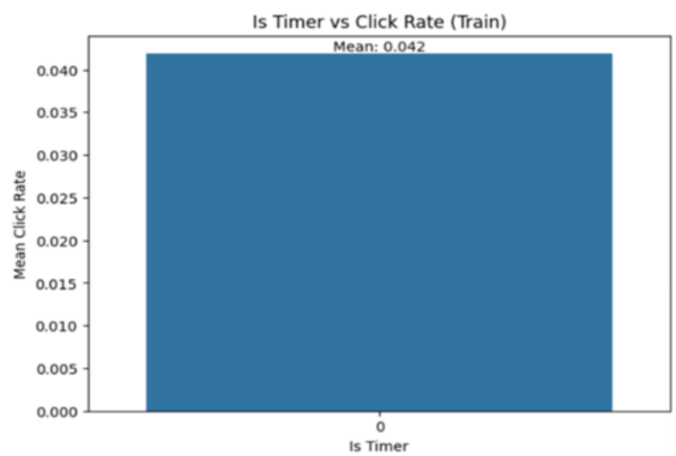


Fig. 33. Is timer vs Click Rate

As we have observed relation between is_personalised variable and click through rate from the visualizations, we are performing t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we expect some relation between the feature and click through rate. We can consider this feature to predict the click through rate.

15) is_quote: In this feature, there are 1888 values in total. There are no missing values in the data. these values lies in the range of 0 to 6. The mean of the values is 0.83 and the standard deviation is 1.03. We have used a bar plot to represent is_quote data and is_quote is present on x-axis and counts of each value is on y-axis. As the range increased the count of each value is decreased from high to low.

We have plotted a bar plot to visualize the relationship between is_quote and click-through rate with is_quote on x-axis and click-through rate on y-axis. The is_quote datapoints lies between 0 to 6 and the click_rate range is in between 0 to 0.05. the highest click_Rate for is_quote is 0.052 at zero. There is a relationship between is_quote and click_Rate.

As we have observed relation between is_quote variable and click through rate from the visualizations, we are performing t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we expect some relation between the feature and click through rate. We can consider this feature to predict the click through rate.

16) is_timer: In this feature, there are 1888 values in total which means we don't have any missing values in the data. this feature is_timer has only one value i.e., 0. The mean and standard deviation for this feature is zero. We have plotted a bar graph to visualize this data.

We have plotted a bar plot to visualize the relationship between is_timer and the click-through rate with is_timer on x-axis and click_Rate on y-axis. And the mean is displayed on 0.042 and there is only one value zero in this is_timer data. There is no relationship between is_timer and click-through rate. So we are not considering this feature for this tasks.

According the data, we can see that the value of is_timer is always zero, which indicates that this feature does not affect the click through rate. We are not considering this feature to predict the click through rate. Also, calculated the correlation

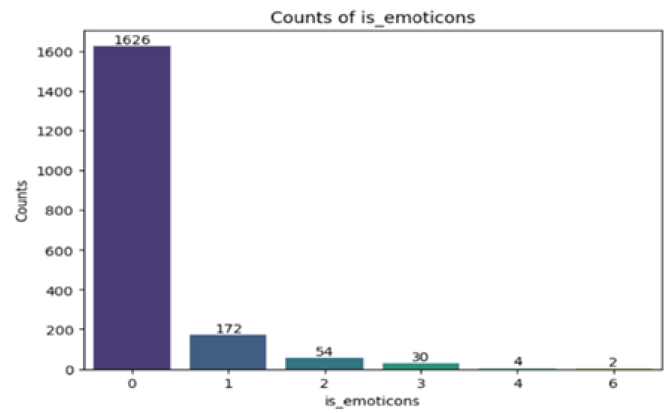


Fig. 34. count of is emoticons

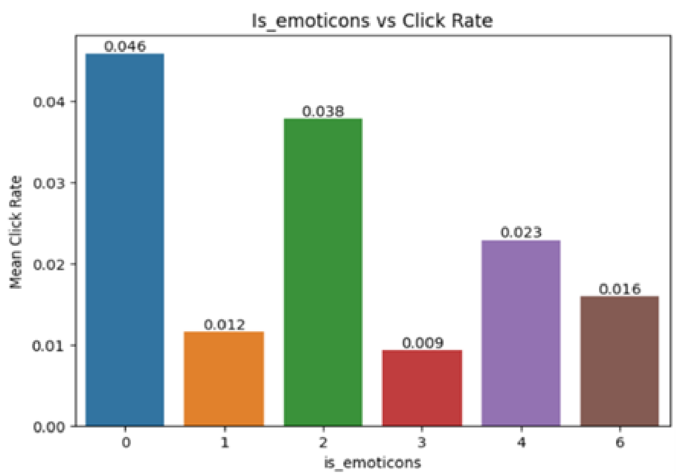


Fig. 35. is emoticons vs click rate

as a proof which is NA.

17) is_emoticons: This feature have 1888 data points in total, that means there is no missing data. These data points lie in the range of 0 to 6. The mean of the values is 0.21 and standard deviation is 0.61. We have used a barplot to visualize this feature. As we can see that the data has more zeroes in it and the data is less at the upper bound.

To visualize the relationship of is_emoticons with respect to click_rate with is_emoticons on x-axis and click_Rate on y-axis. This graph shows that is_emoticons of category zero has more click_rate with 0.046. There may be a relationship between is_emoticons and click-through rate.

As we have observed relation between is_emoticons variable and click through rate from the visualizations, we are performing t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we can consider this feature to predict the click through rate.

18) is_discount: This feature has 1888 values in total and there is no missing values in it. The data lies in only two values they are zero and one. The mean for the data is .04 and the standard deviation is 0.2. We have visualized this data

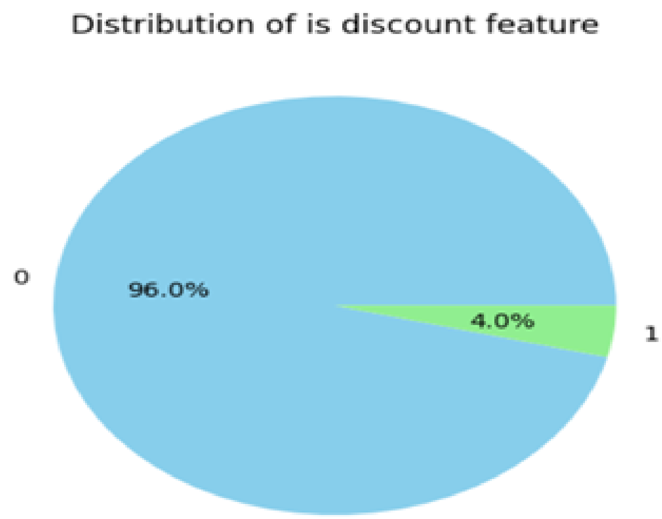


Fig. 36. Distribution of is discount feature

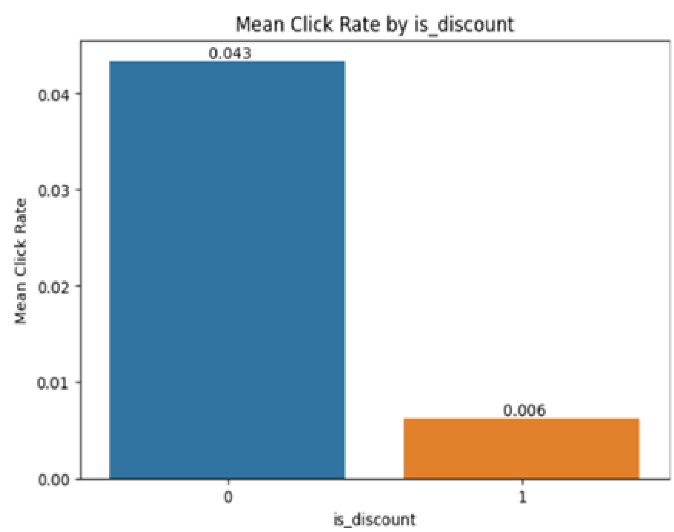


Fig. 37. mean click rate by is discount

with a pie chart. 96% of the data is 0's, only 4% of data points are 1's.

To know the relationship between is_discount and click_rate we have plotted a bar plot. From the plot we can see that the mean click through rate when is_discount is 0 is 0.043 and when 1 is 0.006.

We are performing a t-test to understand the relation between is_discount and click through rate. P-value is 0.65 which indicates that we do not have enough evidence to reject the null hypothesis. It indicates that there is no relationship between this feature and click through rate. Thus, we are not considering is_discount feature to predict click through rate.

19) is_price: This feature contains 1888 data points. This data does not have any missing points. These values lies in the range of 0 to 14999. The mean for this data is 40.2 and the standard deviation is 554. We have used a bar plot to visualize the data points. Most of the values are 0's and remaining values

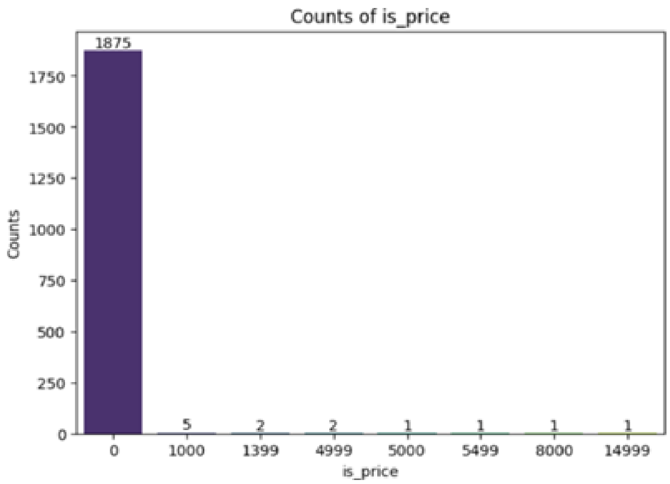


Fig. 38. Count of is price

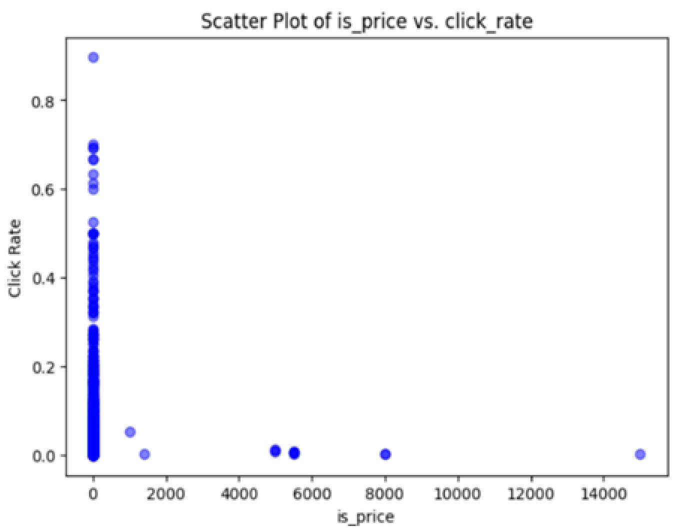


Fig. 39. is price vs click rate

are less in counts.

We are plotting a scatter plot to understand the relation between is_price variable and click-through rate. We can observe that the clicks and the feature are related near to the origin.

As we have observed relation between is_urgency variable and click through rate from the visualizations, we are performing t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we can consider this feature to predict the click through rate.

20) is_urgency: This feature has 1888 datapoints in total that shows that there are no missing values in the data. These values lies in between 0 and 1. Mean is 0.11 and standard deviation is 0.32. we have used a pie chart to visualize the data. In this there are 88.8% of zeroes and 11.2% of ones'. This shows that there is not much urgency.

To understand the relation between click through rate and

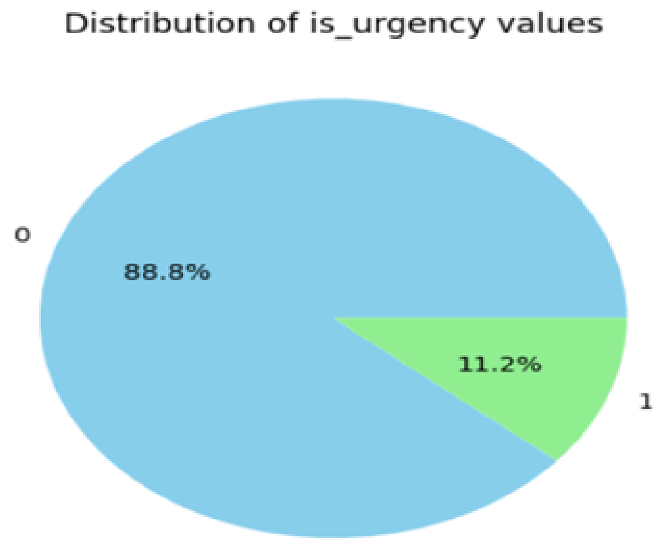


Fig. 40. Distribution of is urgency value

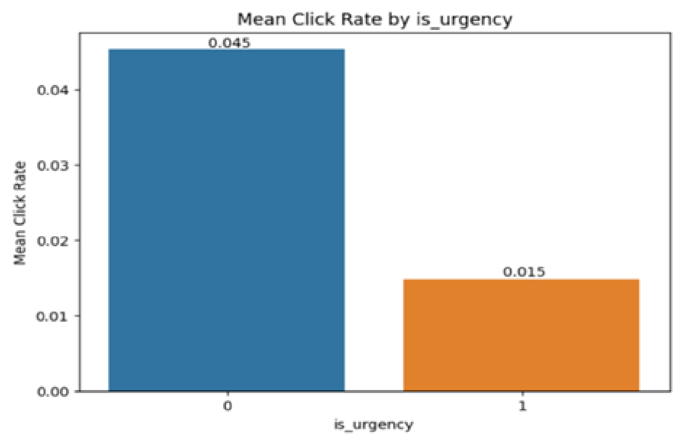


Fig. 41. Mean Click Rate by is_urgency

is_urgency, we are plotting a bar plot. When, the email is sent as an urgent email, the mean click rate is lower than when it is not urgent.

As we have observed the relation between is_urgency variable and click-through rate from the visualizations, we are performing t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we can consider this feature to predict the click-through rate.

21) target_audience: This feature target audience has 1888 data points which lie in the range of 0 to 16 and there are no missing values in it with a mean of 11.6 and a standard deviation is 2.95. we have used a bar lot to visualize this feature. And most of the target audience are under category 12. And the count is 1169. The least number of audience came under the category 0 with a count of 3.

To understand the relation between click through rate and target audience, we are plotting a bar graph. We can observe that the mean value of click through rate is highest when the

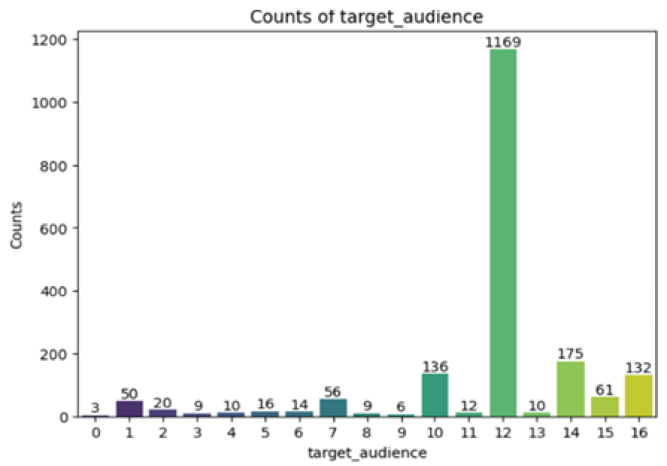


Fig. 42. Count of Target Audience

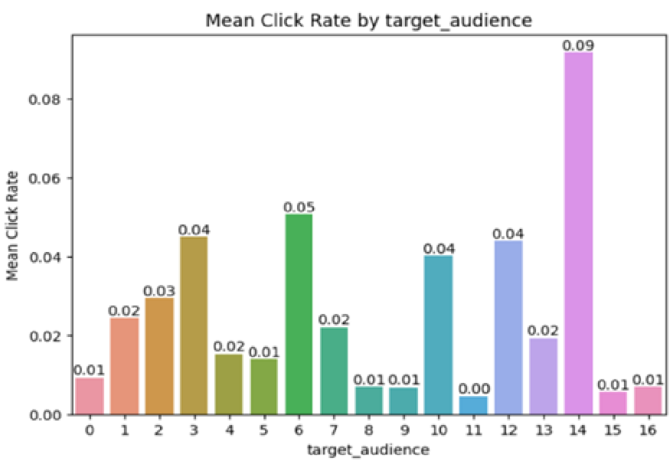


Fig. 43. Mean click rate by target audience

target audience is 14.

As we have observed relation between target audience and click through rate from the visualizations, we are performing t-test and OLS linear regression model test to get the statistical evidence. The p-value is very low, which indicates that the null hypothesis should be rejected. Hence, we can consider this feature to predict the click through rate.

VI. Comprehensive Data Analysis

After analyzing various data sets, a multivariate regression was conducted with selected data alongside the 'target_audience' variable. This analysis aimed to measure the impact of the selected data, if found to be an independent variable alongside 'target_audience', on the dependent variable, 'click_rate'. This assessment sought to determine the collective influence exerted by both the selected data and the 'target_audience' on the 'click_rate' variable.

The list of selected data is as follows: 'day_of_week', 'times_of_day', 'subject_len', 'no_of_CTA', 'is_emoticons', 'is_urgency', 'is_discount', and 'category.

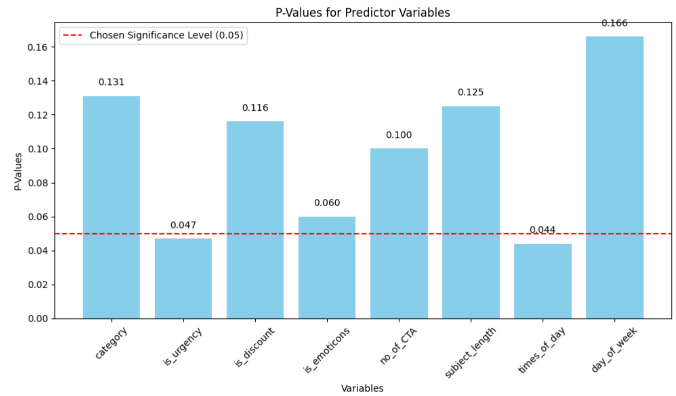


Fig. 44. p-values of 'target_audience' from multivariate regression

Based on the results from the above graph, when analyzed alongside the target audience, it can be inferred that the category of the product, presence of discounts, number of call-to-actions (CTA), length of the subject line, and the day of the week of dispatch do not significantly influence the click rate of the target audience. However, it can be observed that the urgency of the email and the time of day the email is sent within a day significantly influence the click rate when considered alongside the target audience. Furthermore, although the inclusion of emoticons falls slightly short of the selected significance level, with a value of approximately 0.06, it indicates a noteworthy metric that, while not prioritized in terms of importance, warrants consideration in the analysis.

VII. Implementation

A. Data Pre-processing:

Now that we have analyzed the data and selected features that will affect the click-through rate, we are storing these features in a data frame. To prepare the data to be appropriate for considering it as an input to the machine learning models, we have to make sure to remove outliers from the numerical data and change the categorical features to numerical values by using one hot encoding. Using the train_test_split function in the sklearn library, we are splitting the data into two parts. Train data with 70% of the data and test data with 30% of the data.

B. Training Phase

With statistical evidence, we have selected the below mentioned features to be effecting the prediction of click through rate:

- sender
- category
- product
- day_of_week
- is_weekend
- times_of_day
- no_of_CTA
- mean_CTA_len
- is_image

- is_personalised
- is_quote
- is_emoticons
- is_price
- is_urgency
- target_audience
- subject_len
- body_len
- mean_paragraph_len

Considering all the above-mentioned features, we are training a machine learning model, to predict the click-through rate of the advertisement emails. The target variable click-through rate represents the probability of the customer clicking on the advertisement which is a continuous value. This is a regression problem, where we are trying to predict the probability of a click or click-through rate.

Now using the training data, which is 70% of the whole data, we are training a machine learning model considering multiple regression-based models as mentioned below:

- Linear Regression
- K Nearest Neighbours Regressor
- Decision Tree Regressor
- Random Forest Regressor [2]
- XG Boost Regressor [3]
- Gradient Boosting

For enhancement of the performance of the models, applying techniques like Principle Component Analysis and Cross Validation. Also training the data on multiple deep learning models to understand non-linear and high dimensional correlations:

- Convolution Neural Network
- Deep Neural Network
- Feed Forward Neural Network
- Dense Convolution Neural Network
- Deep and Wide Neural Network
- Residual Neural Network
- Long Short-Term Memory Network

C. Test Phase

After training the model, we predicted the click-through rate on the test samples and calculated the mean squared error, R-squared value, and explained the variance score to understand the performance of the model on unseen data.

VIII. Model Performance Analysis

A. Base Machine Learning Models

After training the model using multiple regression-based algorithms, below is the performance of each model. As this is a regression problem, we are considering mean squared error, R-squared value, and explained variance score.

Mean Squared Error: This calculates the average square difference between predicted and actual values. A lower value indicates that the model fits better.

R-squared value: The proportion of variance in the target variable that is predictable from independent variables. A higher value indicates that the model fits better.

Model Name	Mean Squared Error	R-squared	Explained Variance Score	Time Taken (sec)
LinearRegression	0.0056	0.1133	0.1141	0.0346
KNeighborsRegressor	0.0064	-0.0153	-0.0123	0.0081
DecisionTreeRegressor	0.0071	-0.1376	-0.1350	0.0520
RandomForestRegressor	0.0031	0.5065	0.5069	1.1333
XGBRegressor	0.0027	0.5643	0.5653	0.1934
GradientBoostingRegressor	0.0040	0.3621	0.3622	0.2999

Fig. 45. Performance Metrics

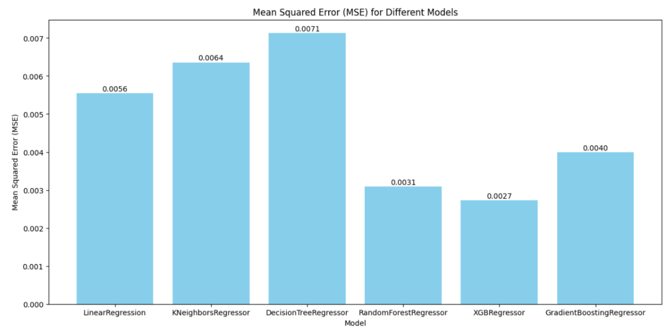


Fig. 46. Mean Square Error

Explained Variance Score: Similar to R-squared but quantified version. A higher value indicates that the model fits better.

B. Machine Learning Models post Principle Component Analysis

The above models are trained with 18 features, among which there are 6 categorical variables and 12 numerical variables. These categorical variables are converted to numerical variables by applying One Hot Encoding. In one hot encoding, each categorical variable in divided into multiple columns. For example, the feature times_of_day is having 3 values namely Noon, Morning and Evening. So, 3 columns are formed and the values are 0 or 1 depending on the variable. Considering 6 categorical features, these features will be divided into multiple sparse features. Each of these are considered to be principle components. We have a total of 96 principle components.

By using principle component analysis, we are trying to understand the time taken for training each model and the evaluation of metrics in comparison with the base 6 models trained previously. We are trying to observed if we can see any clear difference.

From the cumulative explained variance, we can see that the cumulative % for 40 principle components is 0.9870. This means that if we consider 40 principle components then 98.70% of the variance can be captured. Training the model with 40 features should be easier than all features.

From the metrics table, we can infer that the time taken to train Linear Regression and K Neighbours Regressor were comparatively very low when using 40 principle components. From the graph, we can see that the performance of Random Forest Regressor, XG Boost Regressor and Gradient Boosting Regressor is higher than the other models, but not higher than the base models. Considering the time taken and mean square

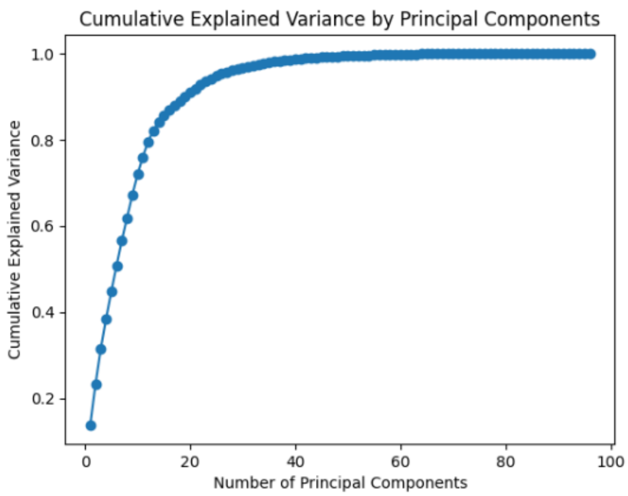


Fig. 47. Cumulative Explained Variance

Model Name	Mean Squared Error	R-squared	Explained Variance Score	Time Taken (sec)
LinearRegression	0.0050	0.2040	0.2041	0.0035
KNeighborsRegressor	0.0064	-0.0154	-0.0123	0.0007
DecisionTreeRegressor	0.0069	-0.0967	-0.0967	0.0942
RandomForestRegressor	0.0034	0.4572	0.4612	6.9317
XGBRegressor	0.0029	0.5365	0.5366	1.8639
GradientBoostingRegressor	0.0034	0.4644	0.4644	1.4076

Fig. 48. Performance metrics after PCA

error values, applying principle component analysis to the data set did not add enough value.

C. Machine Learning Models post Cross Validation

Cross validation is basically a technique of re-sampling the data and dividing them into multiple groups. This step is necessary to ensure that the data used for training and validation purposes are randomly shuffled. This ensures to increase the generalizability of the model, in turn its metrics.

The model is trained multiple times on different combination of data groups, for example in our case we are taking cross validation with 8 folds, which means that the whole train data is divided into 8 parts. In the first iteration, first 7 parts are considered for training and remaining one part for validation. This process is repeated 8 times as we have 8 folds, and the

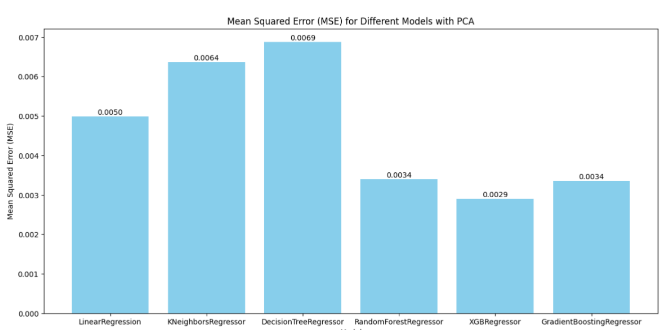


Fig. 49. MSE post PCA

Model Name	Mean Squared Error	R-squared	Explained Variance Score	Time Taken (sec)
LinearRegression	0.0046	0.1651	0.1726	0.2109
KNeighborsRegressor	0.0050	0.0786	0.0860	0.1237
DecisionTreeRegressor	0.0055	-0.0446	0.0319	0.1852
RandomForestRegressor	0.0027	0.5072	0.5116	8.6296
XGBRegressor	0.0027	0.5092	0.5122	2.1016
GradientBoostingRegressor	0.0032	0.4218	0.4273	1.9983

Fig. 50. Performance Metrics post Cross Validation

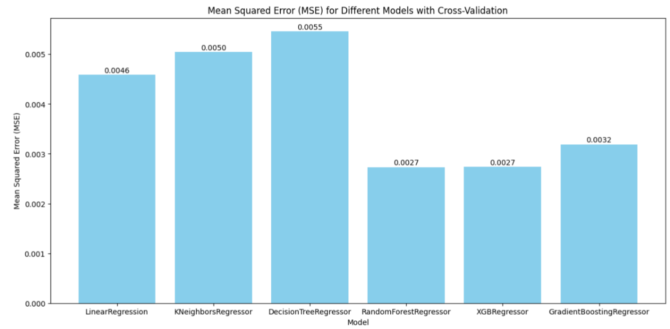


Fig. 51. MSE after Cross Validation

evaluation metrics like mean square error is calculated. To get the total mean square error, we are averaging all the MSE scores from 8 iterations.

From the output, we can see that the mean square error of Random Forest Regressor, XG Boost Regressor and Gradient Boosting Regressor are less than the other models. We can see a clear decrease in the mean square error for Linear Regression Model, K Neighbours Regressor, Random Forest Regressor and Gradient Boosting Regressor after using cross validation techniques. Though the mean square error has decreased, the base model of XG Boost Regressor is having the least mean square error in terms of decimals.

D. Deep Learning Models

In order to capture complex relationships between features and their hierarchical representations, deep neural networks were used to train the data. Models with multiple layers will be able to capture complex non-linear relationships between the features. Though we are dealing with structured data, there can be non-linear relationships between features which cannot be captured by machine learning models, thus training the data on deep learning models to check for performance improvements and model generalizability.

Starting with the basic convolutional neural network with 4 layers. The first layer is a 1D convolutional layer with 32 filters each of size 3 and ReLU activation function to introduce non-linearity. Next is the flatten layer which is used to flatten the output from the previous layer and to prepare the data for next layer. Next, we have 2 dense layers which has fully connected layers with 64 and 1 neuron respectively. The last layer acts as a output layer where we have one neuron to deal with regression problems. The model was trained using scaled and reshaped train data and then tested on unseen data to check the generalizability. We are using an Adam optimizer and the loss type considered is mean squared error. The mean square

Deep Leaning Model	Mean Squared Error	R-squared	Explained Variance Score
Convolution Neural Network	0.004961721	0.207767506	0.20814479
Deep Neural Network	0.004521347	0.278081441	0.296951976
Feed Forward Neural Network	0.00383325	0.38794916	0.400764408
Dense Convolution Neural Network	0.00398078	0.36439319	0.365013938
Deep and Wide Neural Network	0.032697889	-4.22083603	-4.053989379
Residual Neural Network	0.005094074	0.186634731	0.209481821
Long Short-Term Memory Network	0.006033161	0.036691743	0.037345741

Fig. 52. Performance Metrics of Deep Learning Models

Feature	Description	Performed By	Contribution
Data Understanding	Understanding the meaning of each feature, identifying the data type and the values in each feature	Chae	6%
Exploratory Data Analysis	Univariate Analysis for Sender, category, product	Jyothi	3%
	Univariate Analysis for day_of_week, is_weekend, times_of_day	Chae	3%
	Univariate Analysis for subject_len, body_len, mean_paragraph_len	Keerthi	3%
	Univariate Analysis for no_of_CTA, mean_CTA_len, is_image, is_personalised, is_quote, is_timer	Sajid	3%
	Univariate Analysis for is_emoticons, is_discount, is_price, is_urgency, target_audience	Siri	3%
Correlation between features and target	Analysing the correlation matrix between all features and target	Sajid	6%
Data Visualization	Visualization to understand relation between features: Sender, category, product, and target	Jyothi	3%
	Visualization to understand relation between features: day_of_week, is_weekend, times_of_day, and target	Chae	3%
	Visualization to understand relation between features: subject_len, body_len, mean_paragraph_len, and target	Keerthi	3%
	Visualization to understand relation between features: no_of_CTA, mean_CTA_len, is_image, is_personalised, is_quote, is_timer, and target	Sajid	3%
	Visualization to understand relation between features: is_emoticons, is_discount, is_price, is_urgency, target_audience, and target	Siri	3%
Statistical Analysis	Statistical test to prove relation between features: Sender, category, product, and target	Jyothi	3%
	Statistical test to prove relation between features: day_of_week, is_weekend, times_of_day, and target	Chae	3%
	Statistical test to prove relation between features: subject_len, body_len, mean_paragraph_len, and target	Keerthi	3%
	Statistical test to prove relation between features: no_of_CTA, mean_CTA_len, is_image, is_personalised, is_quote, is_timer, and target	Sajid	3%
	Statistical test to prove relation between features: is_emoticons, is_discount, is_price, is_urgency, target_audience, and target	Siri	3%
Data Distribution of Target Variable	Data distribution of target variable using Bootstrap sampling method	Siri	6%
Data Preprocessing	Converting categorical variable to numerical variables, Remove Outliers and Normalize values in numerical features	Jyothi	6%
Comprehensive Data Analysis	Explore the scope of expanding dataset and Identifying seasonal trends	Chae	5%
Model Training: Machine Learning Models	Training data on multiple machine learning models and boosting models and analysing the performance on test data	Keerthi	6%
Principle Component Analysis	Performing PCA, training data on multiple machine learning models with selected principle components and analysing the performance on test data	Sajid	5%
Model Training post Cross Validation	Splitting train and test data using cross validation techniques and training data for multiple iterations	Siri	5%
Model Training: Deep Learning Models	Training data on multiple simple convolutional neural network models and analysing the performance on test data	Jyothi	5%
	Training data on multiple deep complex neural networks and analysing the performance on test data. Training the best model on cross validation data.	Keerthi	5%

Fig. 53. Work Completed

error for the model on test data is 0.0038. Though, the mean square error is quite less, when compared to machine learning models, XG Boosting Regressor having lesser mean squared error.

Next we are using the Deep Convolutional Network with 8 layers. The first three layers are 1D convolutional layers with 64, 128 and 256 filters each of size 3 respectively for each layer and ReLU activation function to introduce non-linearity. Next is the global average pooling layer which performs global average pooling i.e., reducing the spatial dimensions of input data. Next we have 3 dense layers which are fully connected layers with 128, 64 and 1 neuron respectively. The last layer acts as a output layer where we have one neuron to deal with regression problems. We have a dropout layer after the first dense layer to prevent over-fitting of the model. The model was trained using scaled and reshaped train data and then tested on unseen data to check the generalizability. We are using an Adam optimizer with 0.001 learning rate and the loss type considered is mean squared error. The mean square error for the model on test data is 0.0044. The mean square error is higher than the base XG Boosting Model.

Next, we are using a feed forward neural network, 9 layers, where dense and dropout layers are alternatively aligned. The number of neurons are 128, 256, 128, 64 and 1 respectively. Drop out layers are used as we have dense layers which can decrease the generalizability of the model. The MSE of this model is around 0.0041.

Next, we used a dense CNN model with 9 layers. Starting with the input layer which is a convolutional layer with 64 filters. Then, max pooling layer which is responsible reducing the special dimensions of input data. The next layer is again a convolutional layer with same configurations are the first layer. A Flatten layer to flatten the output from the previous layer and to prepare the data for next layer. Next, we have 3 dense layers which has fully connected layers with 128, 64 and 1 neuron respectively. Adding two drop out layers to avoid over-fitting of the model. For this model, the mean square error is 0.0032 which is the least among all the deep neural networks.

We have further trained the model on Deep and Wide Neural network as well as Res-Net Model. The mean square errors of the models are 0.043 and 0.0047 respectively. We have further performed cross validation on Deep CNN model i.e., the model with least MSE to check for performance enhancement. There was a slight increase (0.0035) in the mean squared error which means that the performance of the model has not increased.

IX. Conclusion

Mean Squared Error is minimum when the data is trained on XG Boosting Regressor Model with n estimators as 100 on features selected by performing statistical analysis.

X. Implementation status report

A. *Work completed:*

Refer Figure 53 to check individual contribution

References

[1] GitHub Repository, https://github.com/lakshmi-keerthi/ProjectEA
[2] Facebook Click Predictions, https://quinonero.net/Publications/predicting-clicks-facebook.pdf
[3] Kaggle Dataset, https://www.kaggle.com/datasets/sk4467/email-ctr-prediction?select=train_data.csv (accessed Nov. 19, 2023).
[4] https://www.stat.berkeley.edu/ breiman/randomforest2001.pdf
[5] XG Boost Regressor,, https://machinelearningmastery.com/xgboost-for-regression/
[6] https://research.google.com/pubs/archive/45530.pdf
[7] https://pytorch.org/tutorials/recipes/recipes/defininganeuralnetwork.html
[8] Dense Networks, https://amaarora.github.io/posts/2020-08-02-densenets.html
[9] ResNet, https://towardsdatascience.com/residual-networks-resnets-cb474c7c834a
[10] LSTM Networks, https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html
[11] LSTM Time Series, https://machinelearningmastery.com/lstm-for-time-series-prediction-in-pytorch/
[12] Feed Forward Neural Networks, https://towardsdatascience.com/feed-forward-neural-networks-how-to-successfully-build-them-in-python-
[13] Cross Validation, https://www.researchgate.net/publication/324701535$_{Cross-Validation}$