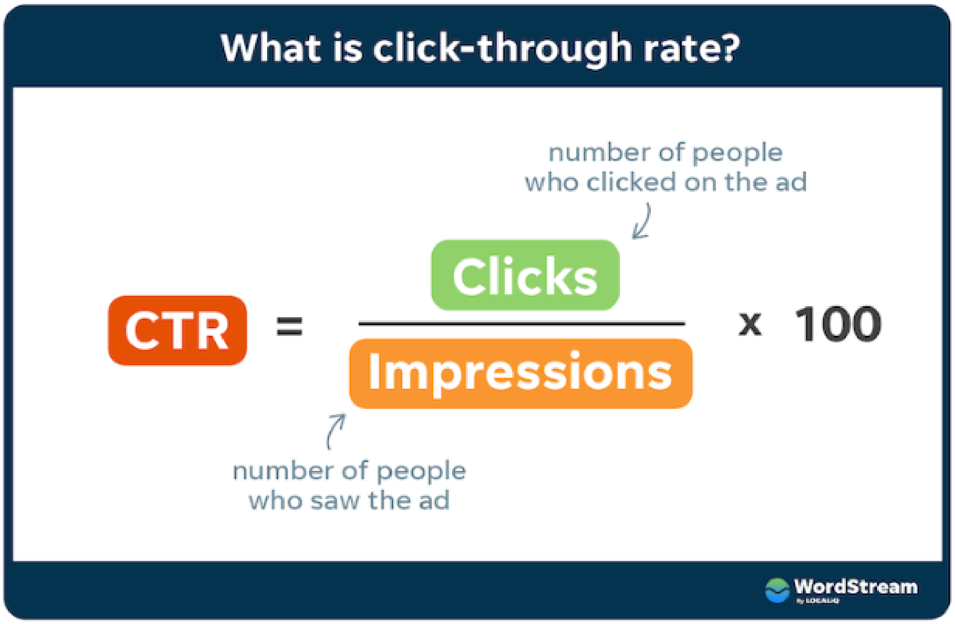
Predicting Click-Through Rate (CTR) of Email Campaigns

Introduction

Companies and industries that are interested in understanding their customers may create an email campaign to gather customer feedback, "An email campaign is a sequence of marketing efforts that contacts multiple recipients at once. Email campaigns are designed to reach out to subscribers at the best time and provide valuable content and relevant offers." (SendPulse, 2022).

Within an email campaign there are embedded calls to action (CTA), which can include things like requesting that the end user reboot their device or complete a survey on the quality of the device remediation (e.g., how the device is performing after a BIOS update or Antivirus enablement). To measure the success of the email campaign, Click Through Rate (CTR) is used, where the higher the CTR the better the email campaign is. See below for how Click Through Rate is calculated (where in this email campaign scenario, Clicks are the number of users who clicked on at least one CTA in the email, and Impressions are the total number of users the email was delivered to), "The formula for CTR looks like this: (Total Clicks on Ad) / (Total Impressions) = Click-Through Rate... A high CTR means that a high percentage of people who see your ad click it.... So a good Google Ads click-through rate is 6-7%+." (Wordstream, 2022). Once calculated, what is a good click-through rate? "A good email click-through rate will vary by industry and type of email campaign, but on average a good click-through rate is about 2.5%." (Forest, 2022).


What is click-through rate?
number of people who clicked on the ad
CTR
=
Clicks
Impressions
x 100
number of people who saw the ad
WordStream

(Wordstream, 2022).

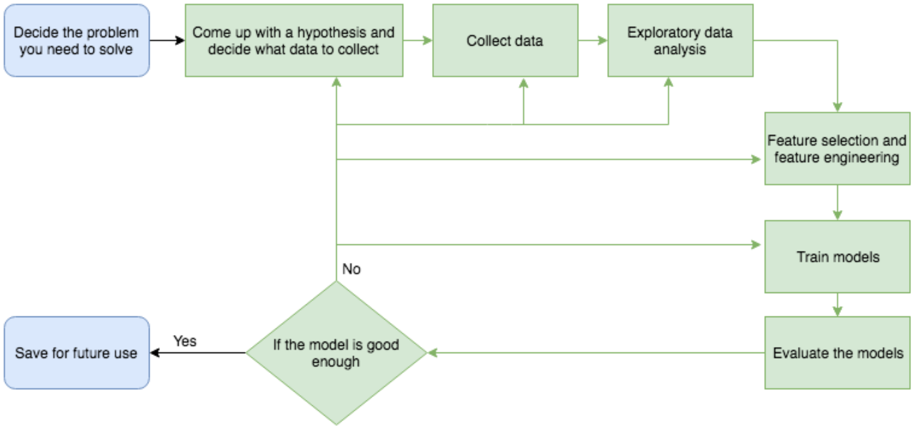
To best predict click-through rate (CTR) of calls to action (CTA) in an email campaign, a dataset was obtained from Kaggle.com (https://www.kaggle.com/datasets/gauravduttakiit/jobathon-august-2022) which has a large volume of records (1,888) and 22 columns or features. The data dictionary can be found in the appendix section.

Methods

Multiple regression models with an 80% train and 20% test sets were used to determine predict the CTR of email campaigns CTA links. Because the Target Variable (click_rate) is a continuous variable, in this research initiative compared the R^2 value for Linear Regression, Lasso

Regression, Ridge Regression, K Neighbors Regressor, Decision Tree Regressor and XGBoost Regressor for model selection. This was a supervised learning model where "the supervisor is the target variable, a column in the data representing values to predict from other columns in the data. The target variable is chosen to represent the answer to a question the organization would like to answer or a value unknown at the time the model is used that would help in decisions. Sometimes supervised learning is also called predictive modeling. The primary predictive modeling algorithms are classification for categorical target variables or regression for continuous target variables." (Abbott, 2014, p. 5).

In addition to the regression models with 1 target and all features, another regression model was used with the 5 best features only derived with highest chi-squared statistics. See below for the process used to evaluate model success:


Decide the problem you need to solve
Come up with a hypothesis and decide what data to collect
Collect data
Exploratory data analysis
Feature selection and feature engineering
Train models
Evaluate the models
If the model is good enough
No
Yes
Save for future use

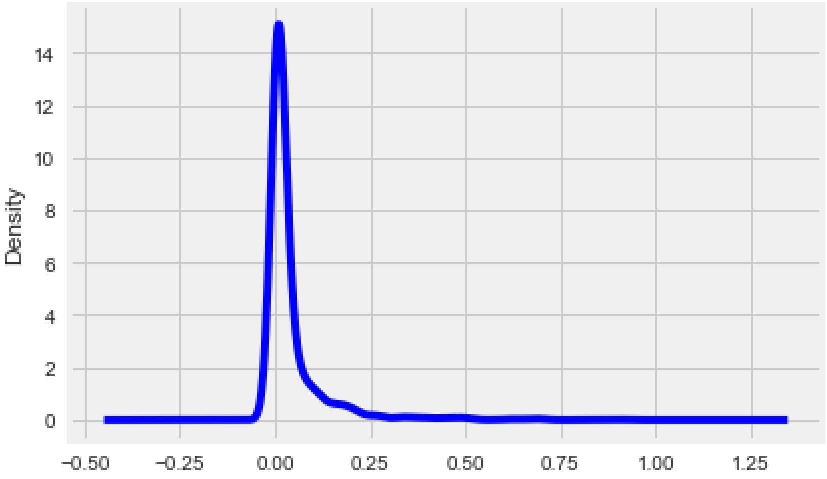
(Keheo, 2018)

To measure the models' performance, the Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 were calculated for the regressor models with all features and 5 best features from chi-squared. The definitions of MSE, MAE, RMSE and R R^2 can be found in the appendix section.

The data was prepared and cleaned to support the creation of several models and evaluation of their performance and results for interpretation (see list below). With these results, a conclusion and recommendations in reference to the original question and problem will be detailed. The categorical feature, times_of_day, was converted into numerical values using label encoder in both the train and test set (dummy variables). A new feature, click_rate_log, was developed as the log of click_rate because click_rate is slightly skewed to the right.

Analysis

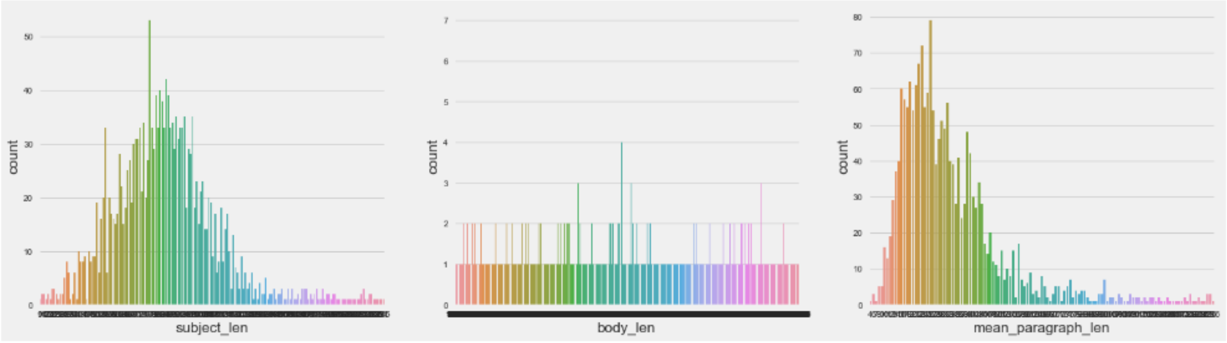
The dataset was explored holistically, and it was found that the number of rows and columns in train dataset was 1888 and 22, respectively and the number of rows and columns in test dataset was 762 and 21, respectively. There is a mix of numerical and categorical data, where the majority are integers and there is no missing data to deal with as there are no null values. The most unique values in the dataset are found in variables 'body_len' and 'click_rate'. Additionally, all features will be useful in this analysis and any categorical features will have dummy variables assigned to them for modeling. To ensure that CTR predictions are accurately gathered, the target variable for the model will be ‘click_rate’. To show the density of the target variable (click_rate), a density plot was used, where it was found that the max click-through rate in the train dataset is 89%, and the train dataset click-through rate is slightly skewed to the right (positive skew).



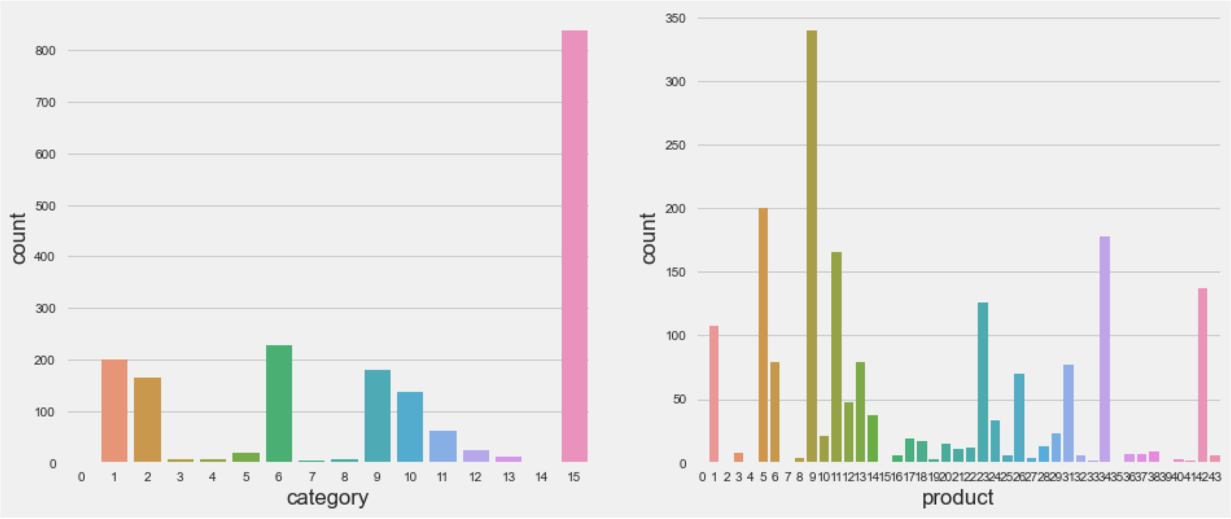
Visualization 1: Target Variable ‘click_rate’ exploration through density chart

Histograms of categorial & numerical data were created to describe the data of single variables (specifically, ‘subject_len’, ‘body_len’, ‘mean_paragraph_len’, ‘category’, ‘product’, ‘no_of_CTA’, ‘mean_CTA_len’, ‘target_audience’, ‘day_of_week’, ‘times_of_day’, ‘is_weekend’, ‘is_image’, ‘is_quote’, ‘is_emoticons’, ‘is_personalised’, ‘is_discount’, ‘is_urgent’, ‘is_timer’) which can be viewed in the below table.

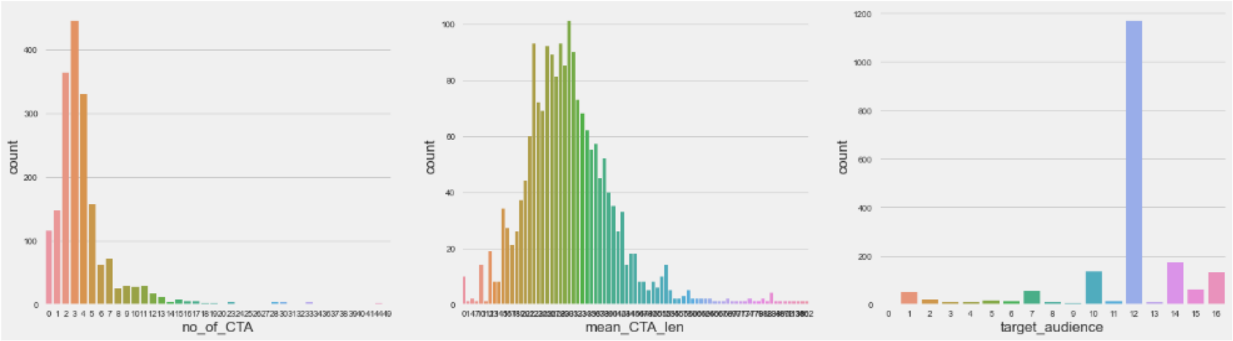
Table 1: Single Variable Histograms



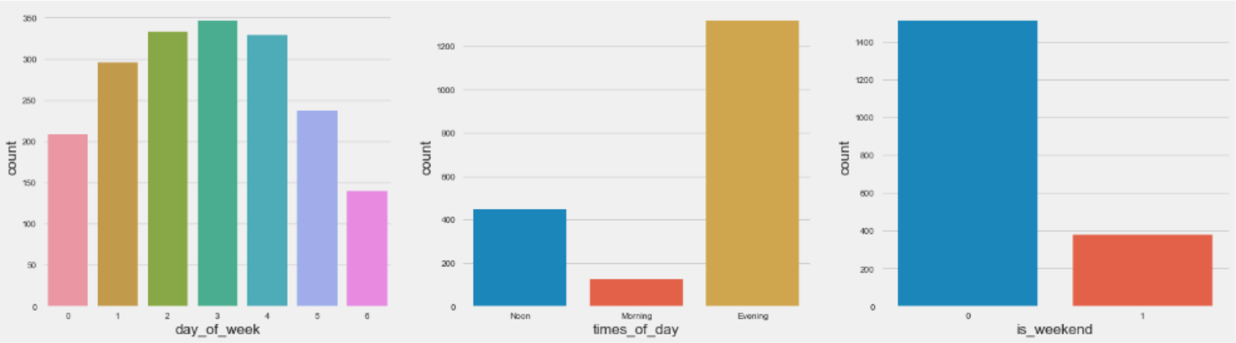
Visualization 2-4: ’subject_len’, ‘body_len’, ‘mean_paragraph_len’



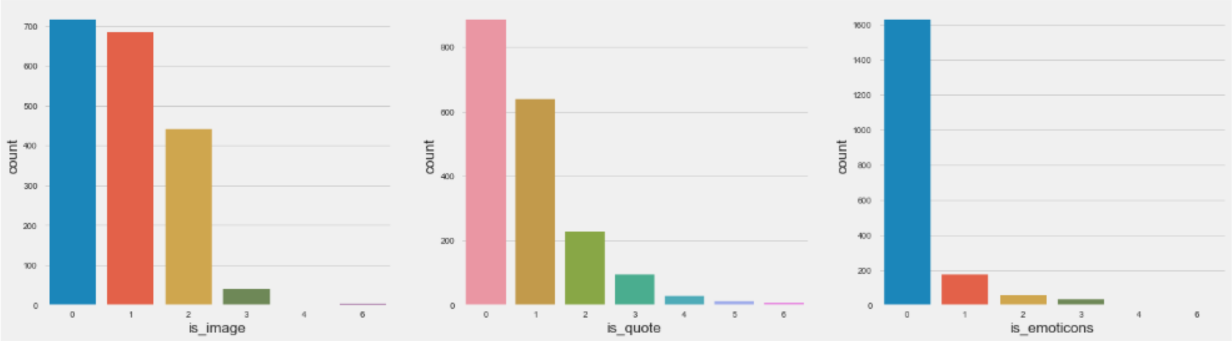
Visualization 5-6: ‘category’, ‘product’



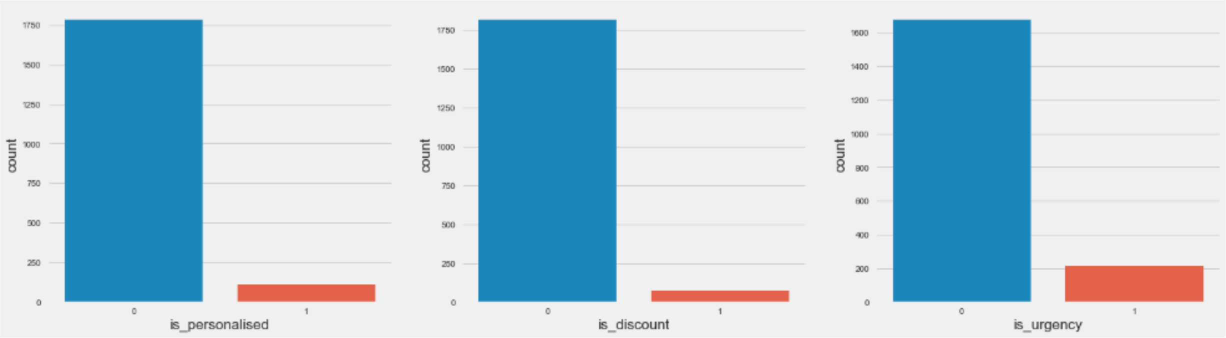
Visualization 7-9: no_of_CTA’, ‘mean_CTA_len’, ‘target_audience’



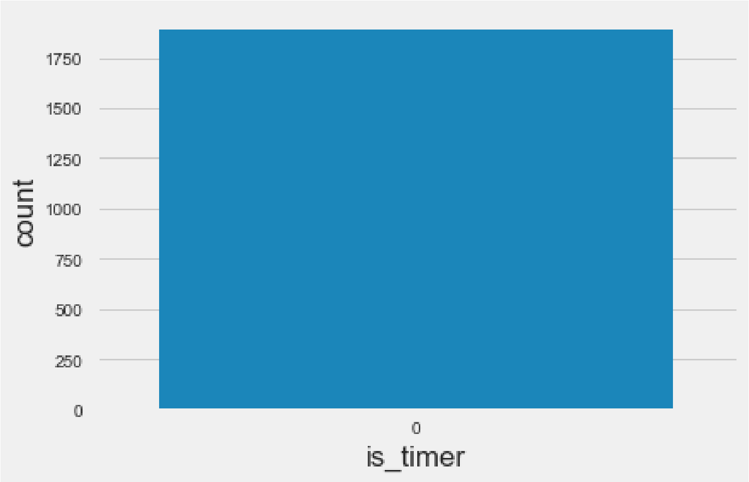
Visualization 10-12: ‘day_of_week’, ‘times_of_day’, ‘is_weekend’



Visualization 13-15: ‘is_image’, ‘is_quote’, ‘is_emoticons’



Visualization 16-18: ‘is_personalised’, ‘is_discount’, ‘is_urgent’

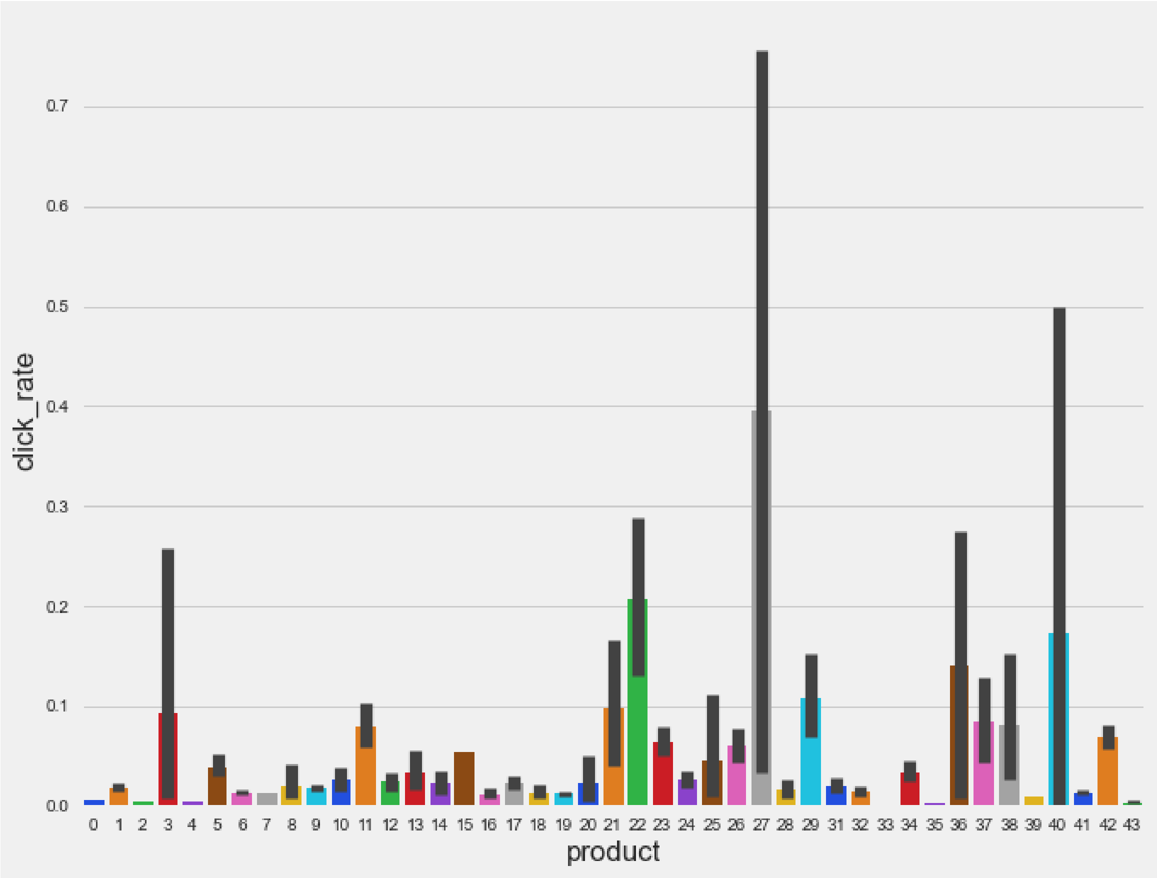


Visualization 19: ’is_timer’

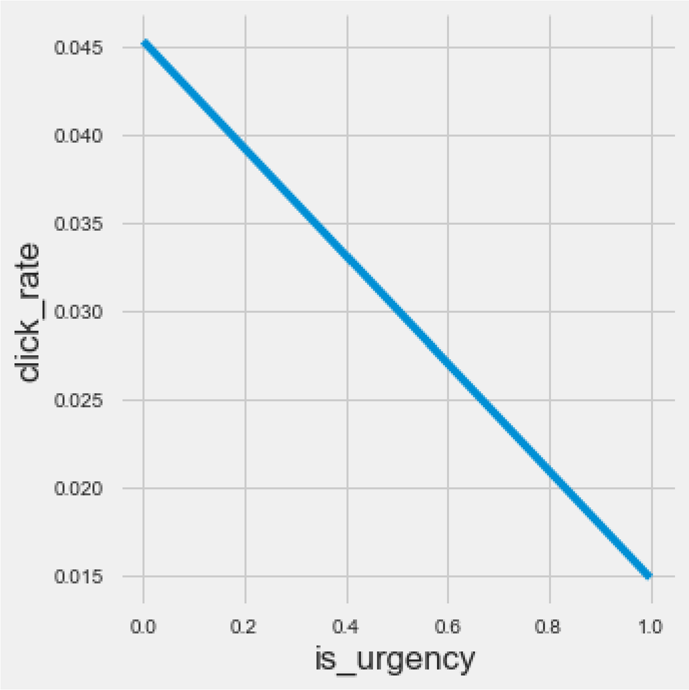
Through these single feature histograms, it was found that the majority of emails have 50-100 characters in the subject, a mean character count of 10-50, an average mean character length between 40-50, related to product category 15 and product types 5, 9 and 34, 0-5 Calls to Action, targeted to audience cluster 12, contained 0-2 images, had 0-3 quotes, contained no emoticons or discounts, personalized to the user, had a timer, and were not urgent. Assuming that 0-6 equals Sunday-Saturday, most of the email campaigns were set on Tuesday-Thursday in the evening.

Bar charts, line charts and Spearman’s correlation matrix of categorial & numerical data were created to describe the data of multiple variables visualized against each other.

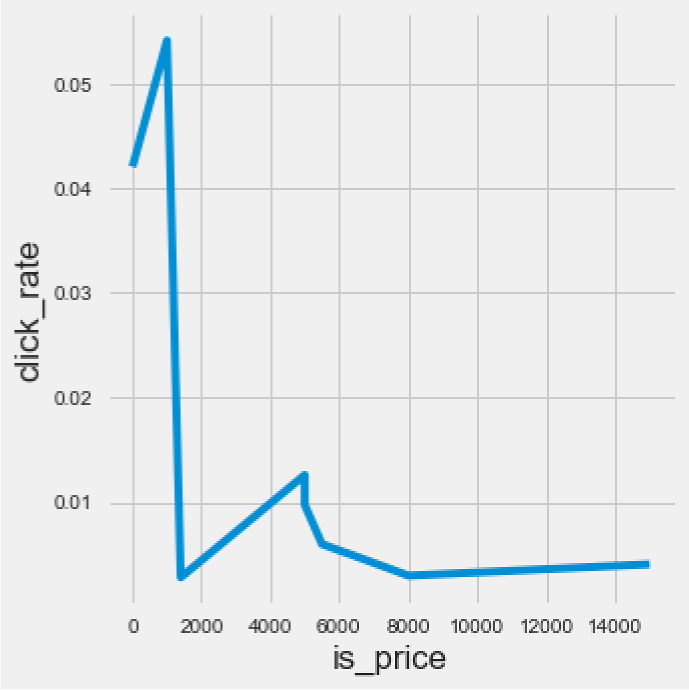
Table 2: Multiple Variable Bar Charts, Line Charts and Spearman’s Correlation Matrix



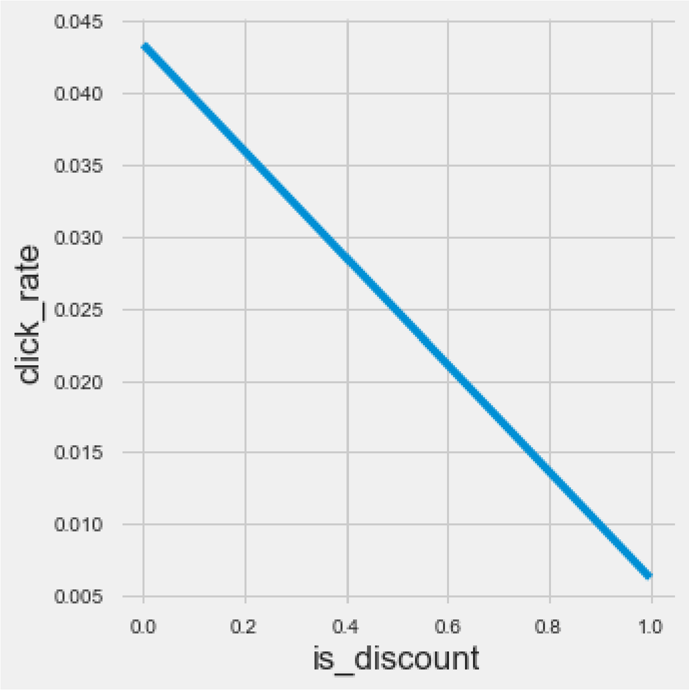
Visualization 20: Bar Chart of ’click_rate’ and ’product’



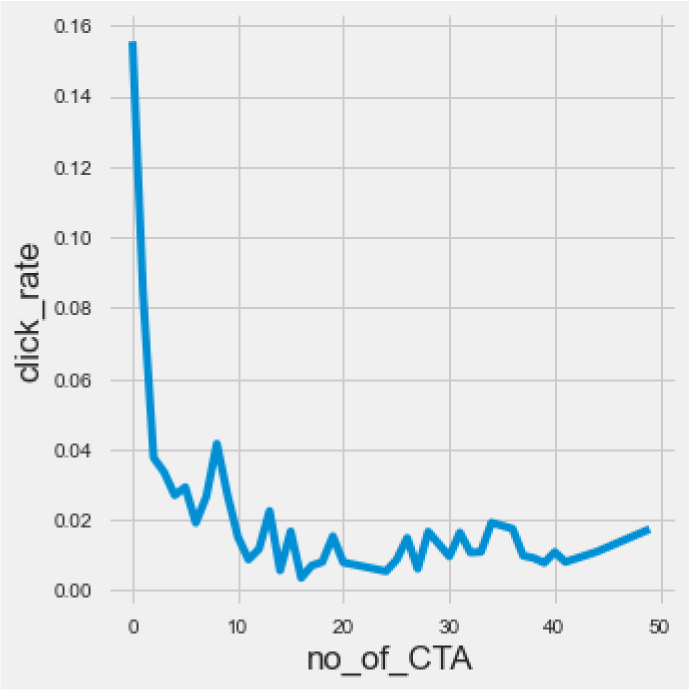
Visualization 21: Line Chart of ’click_rate’ and ’is_urgency’



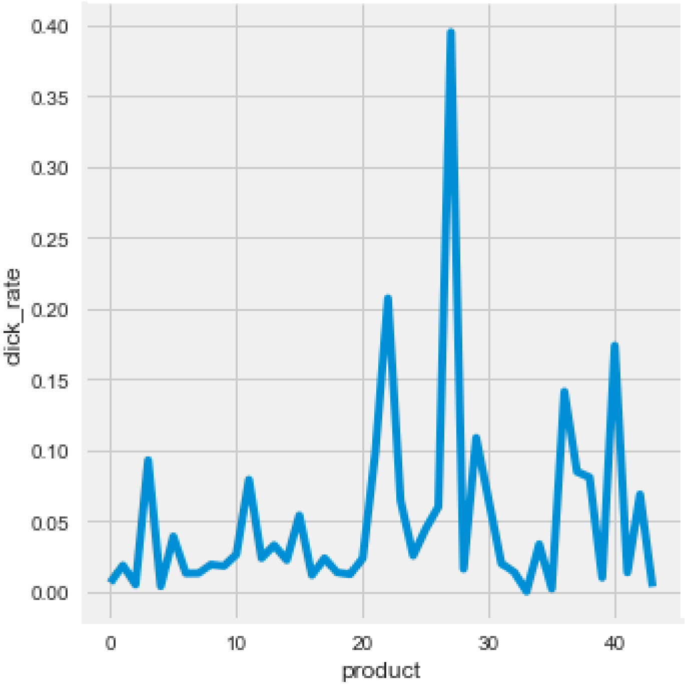
Visualization 22: Line Chart of ’click_rate’ and ’is_price’



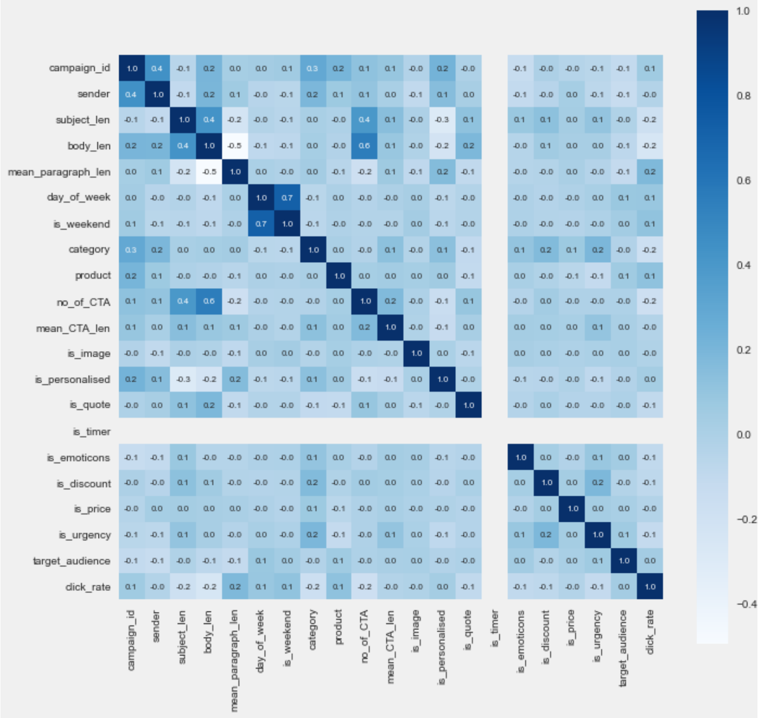
Visualization 23: Line Chart of ’click_rate’ and ’is_discount’



Visualization 24: Line Chart of’click_rate’ and ’no_of_CTA’



Visualization 25: Line Chart of ’click_rate’ and ’product’



Visualization 26: Created to find the features most correlated to the target variable, ‘click_rate’

Through the multiple feature bar chart, it was found that email campaigns related to products 3, 22, 27, 36 and 40 have the highest click-through rate (27 has the highest). Through these multiple feature line charts, it was found that the email campaign is marked as Urgent then the click-through rate is higher (assuming 0 is urgent and 1 is not urgent), the click-through rate is higher if the price is between 0-1500, email campaigns with a discount included have a higher click-through rate (assuming 0 means email includes discount and 1 means does not include discount), email campaigns related to Product 27 has a higher click-through rate, and the click-through rate is much higher for those email campaigns between 0-5 Call to Actions (CTA), so we can assume the higher the CTA the better the email campaign. Through the Spearman’s correlation matrix it was found that the features most positively correlated to the target variable 'click_rate' are ‘mean_paragraph_length’, ‘day_of_week’, ‘is_weekend’, and ‘product’. The features most negatively correlated to the target variable 'click_rate' are ‘subject_len’, ‘body_len’, ‘category’, and ‘no_of_CTA’ (descriptions of these variables can be found in the appendix section under Data Dictionary).

To find the best features of the dataset that are most dependent on the target variable, the Chi-Squared function was used, "Let’s consider a scenario where we need to determine the relationship between the independent category feature (predictor) and dependent category feature(response). In feature selection, we aim to select the features which are highly dependent on the response. When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, the higher the Chi-Square value the feature is more dependent on the response, and it can be selected for model training." (Gajawada, 2022). It was found that the Chi-Squared 5 Best Features are ‘is_emoticons’, ‘is_discount’, ‘is_price’, ‘is_urgency’, and ‘target_audience’ (Python coding results can be found in the appendix section).

To support the training and testing of the models, the dataset was split data into training (80%) and test (20%) sets with ‘click_rate’ feature (continuous variable) as the target. Because the target variable (‘click_rate’) was skewed, an additional feature was created (‘click_rate_log’) which is the log of the target variable. Separate train and test sets were then created with this new target variable. To find the best 5 features, X2 was used, and additional train & test sets were created using the chi-squared function and all features. It was found that the model with the best R^2 value (the performance metric for linear regression models) is XGBoost Regressor. This model will be used as the model of choice for this dataset. (Python coding results can be found in the appendix section).

Results

A total of 3 XGBoost Regressor models were created to find the model with the best performance using the click-through rate dataset. The XGBoost Regressor models used a combination of all features with the 'click_rate' target variable, all features with the 'click_rate_log' target variable, and the 5 best features from X^2 and the 'click_rate_log' target variable. A summary of the model' results can be found below.

Table 3: XGBoost Model Performance Results

Model	Description	Cross Validation Score	K-Fold CV Average	MSE	RMSE	R2
XGBoost Regressor	All Features with 'click_rate' target variable	0.33	0.43	0.00	0.06	0.56
XGBoost Regressor	All Features with 'click_rate_log' target variable	1.00	1.00	0.00	0.00	1.00
XGBoost Regressor	5 Best Features with 'click_rate_log' target variable	0.08	0.10	0.00	0.07	0.07

Through the model performance results, it was found that the XGBoost Regressor model with all features and the target variable 'click_rate' performed well, with an R^2 of 0.56. This is an acceptable R^2 for real-life scenarios. The XGBoost Regressor model with all features and the target variable 'click_rate_log' performed extremely well with an R^2 of 1.00. The XGBoost Regressor model with the 5 best features from X^2 and the target variable 'click_rate_log' did not perform well, with an R^2 of 0.07.

Conclusion

Multiple XGBoost Regressor models were used on the dataset to predict click-through rate (CTR) of calls to action (CTA) in email campaigns. The XGBoost Regressor model with the target variable transformed to its log and all features performed the best, with an R^2 of 1.00. This will be the model of choice to implement with the Product Team to predict the click-through rate of the email campaigns. Additionally, when creating new email campaigns, the below features will be prioritized as features to focus on to increase click-through rate (CTR).

Table 4: Priority Email Campaign Features

Features	Origin
mean_paragraph_length, day_of_week, is_weekend, and product	Spearman's Correlation Matrix
is_emoticons, is_discount, is_price, is_urgency, and target_audience	Chi-Square (X^2) Test

Assumptions

There were a few assumptions of this research initiative. First, the data provided on the Job-A-Thon Kaggle website is accurate. Second, the Regression model Cross validation, Regression model K-fold CV average, Regression model R2 values, and Regression model MSE and Regression model RMSE scores are all acceptable. It was also assumed that for the ‘is_urgent’ and ‘is_discount’ features, 0 is urgent and 1 is not urgent and 0 means email includes discount, and 1 means does not include discount, respectively. It was also assumed that for the ‘day_of_week’ variable, 0-6 equals Sunday-Saturday.

Limitations

There were some limitations of this research initiative. Primarily, no data of the type of device the email campaign was read from (computer, tablet, mobile phone), date opened, open rate, bounce rate, unsubscribe rate was captured. These features could make the model’s performance better and provide more information to the Product Team on the predicted click-through rate of the email campaign.

Challenges

In terms of challenges, the models selected may not provide a good model performance. If a linear model is appropriate, the histogram should look approximately normal, and the scatterplot of residuals should show random scatter. If we see a curved relationship in the residual plot, the linear model is not appropriate. Another type of residual plot shows the residuals versus the explanatory variable. In that case, alternate models will be used such as a lasso regression, ridge regression, K-neighbors regression, Decision Tree regressor, or XGBoost regressor. Although the Chi-Squared feature extraction was successful, the regression model performance did poorly. Because of this, all features will be used in the regression model.

Future Uses/Additional Applications

Future applications of this regression model will be recommended to use in making Product Team decisions on if the email campaign should be sent out as-is or if any changes need to be made. If after adding the email campaign information into the model and the prediction is 6-7% CTR then the email campaign will be sent out as-is. If the prediction is less than 6% CTR then the email campaign will be modified and run through the model again with the goal of getting a CTR of 6-7% or above.

Recommendations

When advising the Product Team on how to develop a successful email campaign, there are several recommendations to provide. First, the features most positively correlated to click-through rate are email mean paragraph length, day of the week the email is sent out, if the email is sent out on a weekday or weekend, and the type of product the email is referring to. The features most dependent on the target variable (click_rate) are is_emoticons, is_discount, is_price, is_urgency, target_audience. Therefore, when creating email campaigns, it's recommended to focus on these features as a priority as they are positively correlated with click-through rate (CTR). Second, the features most negatively correlated with click-through rate at the length of email subject line (in characters), the length of email body (in characters), and number of calls to action (CTA) in the email. Therefore, when creating email campaigns, it's recommended to focus on these features less and prioritize them lower than the other features. Finally, the XGBoost regression model with click_rate (target variable) transformed to log should be moved forward as the model of choice as the model performance (R2) was very good. While creating new email campaigns and before sending them out, it is advised to enter in the email campaign feature information into the model and review the predicted click-through rate. If the predicted click-through rate is 2.5% or above (as described above as a good click-through rate), the email campaign should be sent out as-is. If the click-through rate is below 2.5%, then the email campaign features should be adjusted starting with the positively correlated features and then rerun through the model to determine the predicted click-through rate.

Implementation Plan

The XGBoost model with click_rate (target variable) transformed to log has been trained and then a testing set used to obtain model performance. The next step would be to do a real-life test with email campaign data including click-through rate (CTR) data from the Product Team. The model isn't necessarily ready for deployment, but it is ready to be beta tested with real-life data from the Product Team. This will ensure that more real-world scenarios that relate or do not relate to click-through rate are captured in the dataset

To improve the model, Product Team members will also be interviewed on email campaigns and click-through rate (CTR) of the call to actions (CTA) and possible methods for improvement. Examples of feedback include different data repositories, new features/measurements from email campaigns (e.g., email campaign scheduled delivery, geographic location of receiver), possible anecdotal features that Product Team members have observed in the field

Ethical Assessment

Ethically, it's important to not emphasize click-through rate so much that it causes employees and online advertisers to lie in order to reach a higher CTR, “Online advertising is facing a new form of challenge – the artificial inflation of click-through rates. We call this practice ‘cyber-rigging’.” (Fisher, 2006). Cyber-rigging is considered “Online advertisers who have entered into pay-per-click advertising arrangements may become the victims of unscrupulous advertising companies, search engines, host website owners, or competitors who target their banner ads. In addition to the potential for financial harm to the advertiser, this activity invalidates the common practice of using click-through rate as a measure of the success and effectiveness of online advertising.” (Fisher, 2006). In analyzing this data, it's important to mention that the findings should not be used to increase cyber-rigging and the Product Team should be made aware of this unethical practice.

Appendix

Regression has different performance metrics than Logistic Regression, and they are:

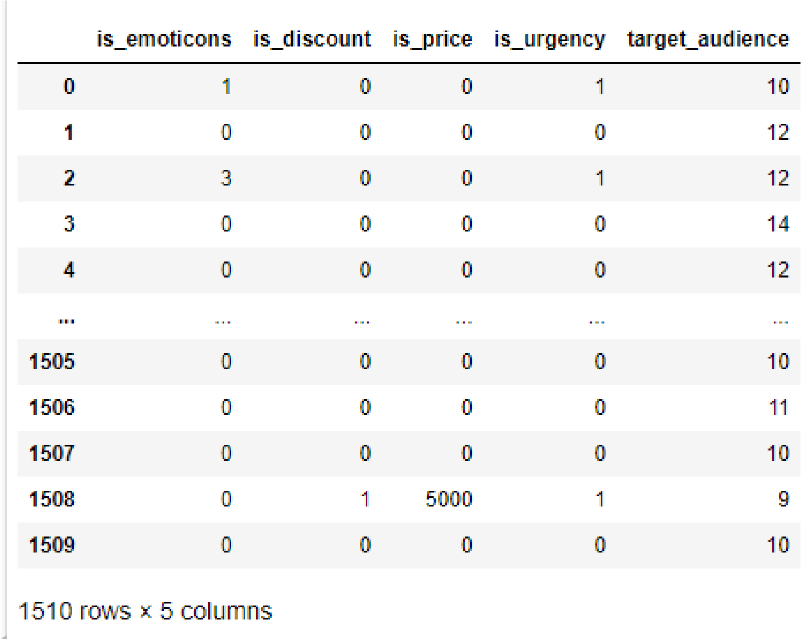
1. “Mean Squared Error (MSE): The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize. MSE penalizes large errors.
2. Mean Absolute Error (MAE): This is simply the average of the absolute difference between the target value and the value predicted by the model. Not preferred in cases where outliers are prominent. MAE does not penalize large errors.
3. Root Mean Squared Error (RMSE): measures the average difference between values predicted by a model and the actual values. RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately
4. R-squared or Coefficient of Determination: This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.” (Yashwanth, 2021)

To understand the definitions of each feature or variable in the dataset, please see the Data Dictionary below.

Data Dictionary	
Feature Name	Feature Description
campaign_id	Unique identifier of a campaign
sender	Sender of an e-mail
subject_len	No. of characters in a subject
body_len	No. of characters in an email body
mean_paragraph_len	Average no. of characters in paragraph of an email

day_of_week	Day on which email is sent
is_weekend	Boolean flag indicating if an email is sent on weekend or not
times_of_day	Times of day when email is sent: Morning, Noon, Evening
category	Category of the product an email is related to
product	Type of the product an email is related to
no_of_CTA	No. of Call To Actions in an email
mean_CTA_len	Average no. of characters in a CTA
target_audience	Cluster label of the target audience
is_image	No. of images in an email
is_quote	No. of quotes in an email
is_personalised	Boolean flag indicating if an email is personalized to the user or not
is_discount	Boolean flag indicating if an email contains a discount or not
is_urgency	Boolean flag indicating if an email contains urgency or not
is_timer	Boolean flag indicating if an email contains a timer or not
is_price	Nominal price listed in email body
click_rate	Click Through Rate, (Total Clicks on CTA) / (Total Impressions)

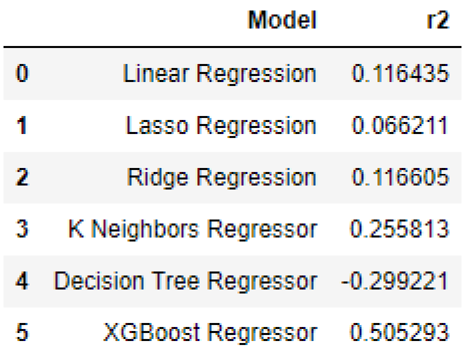
These are the Chi-Squared 5 Best Features found using Python: is_emoticons, is_discount, is_price, is_urgency, target_audience



	is_emoticons	is_discount	is_price	is_urgency	target_audience
0	1	0	0	1	10
1	0	0	0	0	12
2	3	0	0	1	12
3	0	0	0	0	14
4	0	0	0	0	12
...	...	...	...	...	...
1505	0	0	0	0	10
1506	0	0	0	0	11
1507	0	0	0	0	10
1508	0	1	5000	1	9
1509	0	0	0	0	10

1510 rows × 5 columns

The results of the R^2 values of the different models using Python can be found below. Since XGBoost had the highest R^2 value, it was chosen as the model for this initiative.



	Model	r2
0	Linear Regression	0.116435
1	Lasso Regression	0.066211
2	Ridge Regression	0.116605
3	K Neighbors Regressor	0.255813
4	Decision Tree Regressor	-0.299221
5	XGBoost Regressor	0.505293

References

1. Wordstream. (2022, November 17). Click-through rate (CTR): Understanding CTR for PPC. WordStream. Retrieved December 4, 2022, from https://www.wordstream.com/click-through-rate

2. Fisher, J. (2006). Cyber-rigging click-through rates: Exploring the ethical dimensions. Retrieved December 20, 2022, from https://www.researchgate.net/publication/45436601_Cyber-rigging_click-through_rates_Exploring_the_ethical_dimensions
3. Yashwanth, N. V. S. (2021, January 1). Evaluation metrics & model selection in linear regression. Medium. Retrieved December 21, 2022, from https://towardsdatascience.com/evaluation-metrics-model-selection-in-linear-regression-73c7573208be
4. Dhinakaran, A. (2022, December 9). Best practices in ML observability for click-through rate models. Medium. Retrieved December 20, 2022, from https://towardsdatascience.com/best-practices-in-ml-observability-for-click-through-rate-models-8a0c6755a49a
5. Zach. (2020, April 24). What is a good R-squared value? Statology. Retrieved December 20, 2022, from https://www.statology.org/good-r-squared-value/
6. Gajawada, S. K. (2022, September 28). Chi-square test for feature selection in machine learning. Medium. Retrieved December 22, 2022, from https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223
7. Forst, K. (2022, January 7). Improve your click-through rate: 14 powerful tactics. AWeber. Retrieved December 22, 2022, from https://blog.aweber.com/email-marketing/14-powerful-tactics-to-increase-your-email-click-through-rates.htm

8. SendPulse. (2022, December 1). What is an Email Campaign: Definition and Guide. SendPulse. Retrieved December 22, 2022, from https://sendpulse.com/support/glossary/email-campaign