Final Report:
Email Marketing CTR Predictor

Alex Lewis
September 2022

1. Background

In digital marketing, measures of user engagement are crucial indicators of performance. One of the most common engagement metrics is Click Through Rate (CTR) which measures the number of successful engagements divided by the total opportunities provided for users to engage. Advertisers who can successfully model and predict CTRs can tailor their marketing campaigns to build in more of the factors that yield increased CTRs and ultimately improve their marketing success.

In email marketing, the universe of factors marketers must consider when constructing campaigns encompass both content-based ones (how long the email should be, how many Calls-To-Action [CTAs] should be included, etc.) and audience targeting ones (should this email campaign be sent to repeat customers, new customers, or a list of non-customers based on behavior and demographics?). While some of these decisions will be made ahead of time by business necessity, a model that can predict and assign importance to these features can help advertisers build more efficient marketing campaigns and increase their rate of success without wasting resources on costly real-world A/B tests.

My approach to modeling was to treat the click rate as a continuous float to be predicted by regression. I considered multiple regression algorithms ranging from simple to complex and ultimately found an ensemble method to yield the strongest performance as measured by r-squared scoring.

2. Data Wrangling

The data utilized comes from a dataset of email marketing campaigns from Shibu Mohapatra and hosted on Kaggle.com, originally assembled for a hackathon that challenged participants to predict CTR.

I imported the dataset as a Pandas Dataframe and found it to be free of NaN values, however one feature, ‘is_timer’, had only a single binary value of 0 and so was removed. I kept 0 values in my y variable, since a CTR of zero - in other words, an instance where an email campaign failed to earn a single click from a recipient - is an unfortunate reality for many marketers.

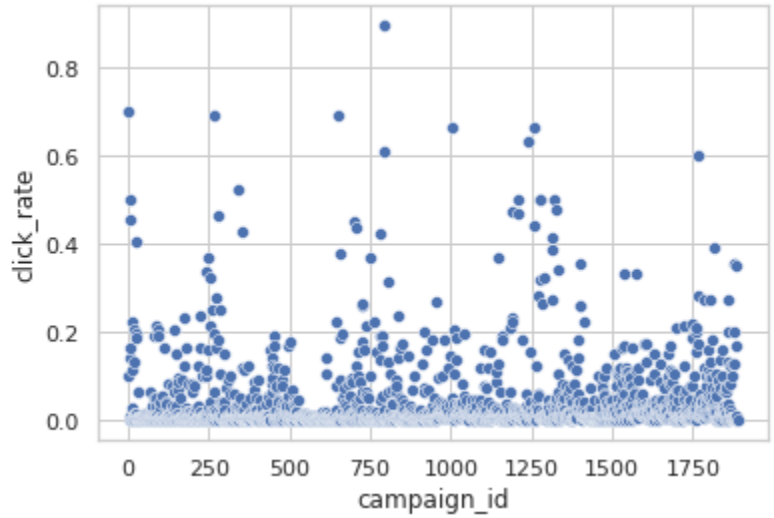
Notably absent in the dataset is any kind of contextual information about two important features, product and audience. In a real-world scenario, the products advertised in an email campaign, and the audiences the campaign is sent to, would factor heavily in strategic business decisions.

However, they may also be proscribed by business leadership as not up for debate. I opted to leave them in my dataset, since ultimately learning about their predictive importance was part of my objective.

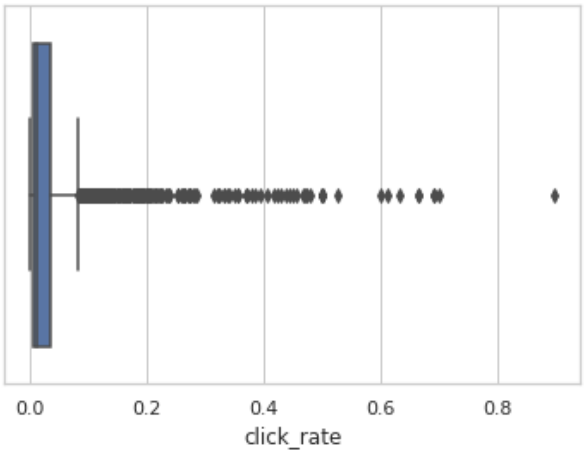
The final shape of my dataset was 1,888 rows x 21 columns prior to encoding categorical variables.

3. Exploratory Data Analysis

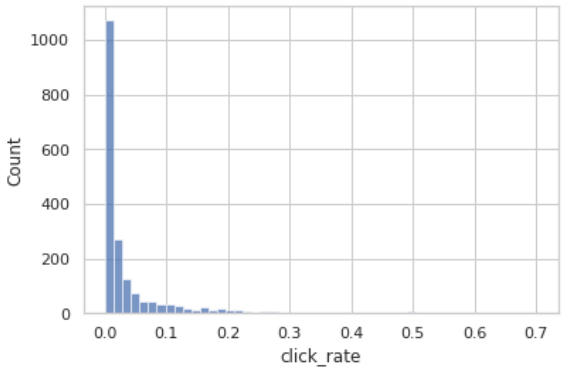
Exploratory analysis of this data set revealed a high degree of variance in the target variable and a significant presence from outliers.

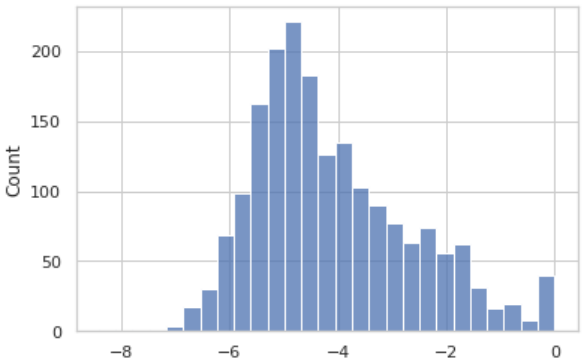


Visualizations revealed most campaigns with CTRs clustered toward the lower end of the scale, with outliers present at values that suggested significant over-performance. From a modeling perspective, this trend provided a helpful hint that an algorithm resistant to outliers, such as Ridge regression, may be a good choice.

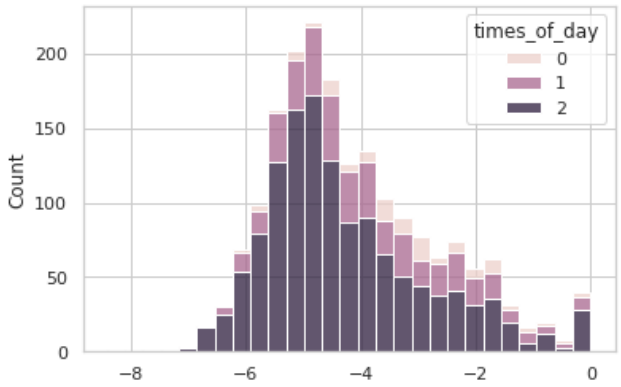


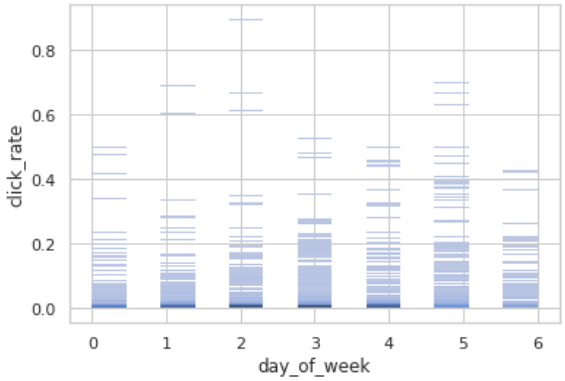
The distribution of CTRs was significantly right-skewed, so I took the log of the variable to get a better look.



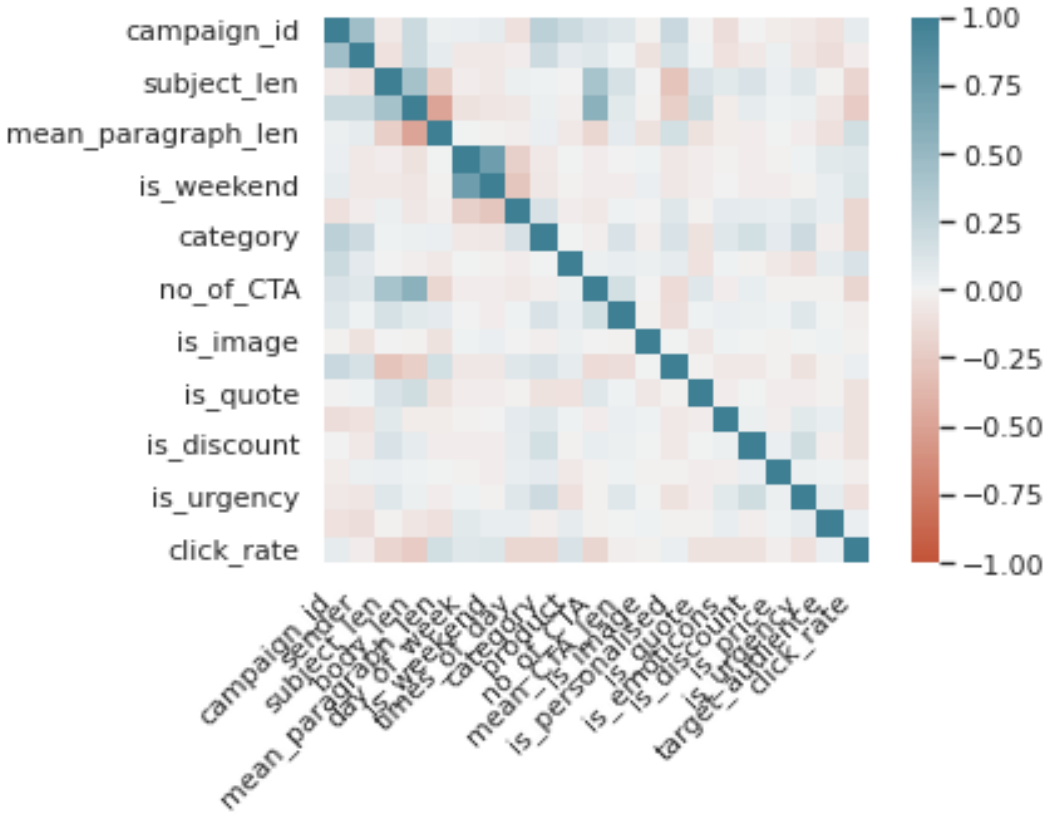


A common focus among marketers is to focus on the timing of email campaigns. Conventional wisdom holds that there are a few universal axioms (morning is better than night; Tuesday and Thursday are the best days of the week) but that industry-specific trends trump general rules. In the case of my dataset, both day and time seemed pretty unimportant to CTR:





After mapping correlation coefficients, it seemed like there might be some inverse correlation between the “length of copy” features: body length and subject line length. There’s some intuitive wisdom here; it’s probably easier for customers to click a CTA if they aren’t blocked by a wall of text.



4. Model Selection & Evaluation

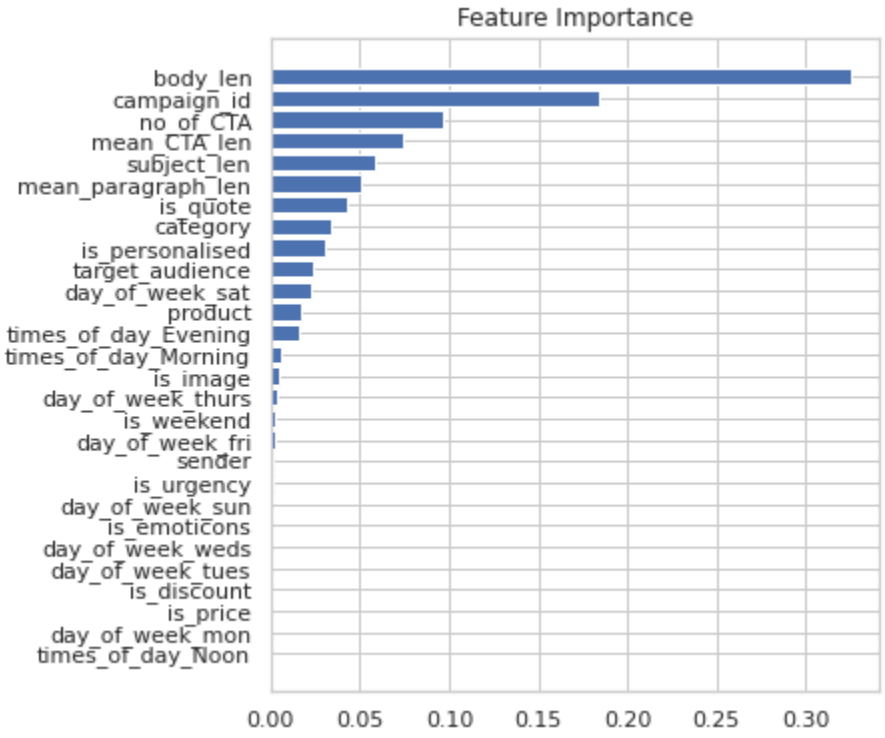
I considered three algorithms - Multiple Linear Regression, KNN Regression, and Gradient Boosting - and weighed their performance based on r-squared score.

Data was split into train and test groups, with the target feature of CTR ('click_rate') set as the Y feature. The dataset was scaled in preprocessing to normalize features. Hyperparameters of the gradient boosting model were adjusted to yield optimum performance.

	model	r2 score
0	linear	0.134361
1	KNN	0.134408
2	GBR	0.483040

While a success rate of less than 50% isn't ideal from a practical standpoint, it ranks among the best on Kaggle for the dataset.

Also important is feature importance; predicting CTRs is of little value to marketers if they don't also know which aspects of campaigns to adjust to yield them. Consistent with what I discovered in EDA, the "length" features rank highly in the GBR model:



5. Recommendations & Improvements

It's likely that the gradient boost algorithm utilized in this analysis could improve its accuracy score after more parameter and hyperparameter tuning.

Since CTRs are usually considered in general terms where a 2.71% isn't thought of as hugely different than a 2.69%, it is probably a good idea to bin the CTR into ranges and attempt modeling as a classification problem. That might be a little more forgiving on accuracy score without losing any actual important information.

Finally, it's clearly a worthy endeavor to attempt additional ensemble methods, given how significantly the GBR outperformed the linear and KNN regressions.