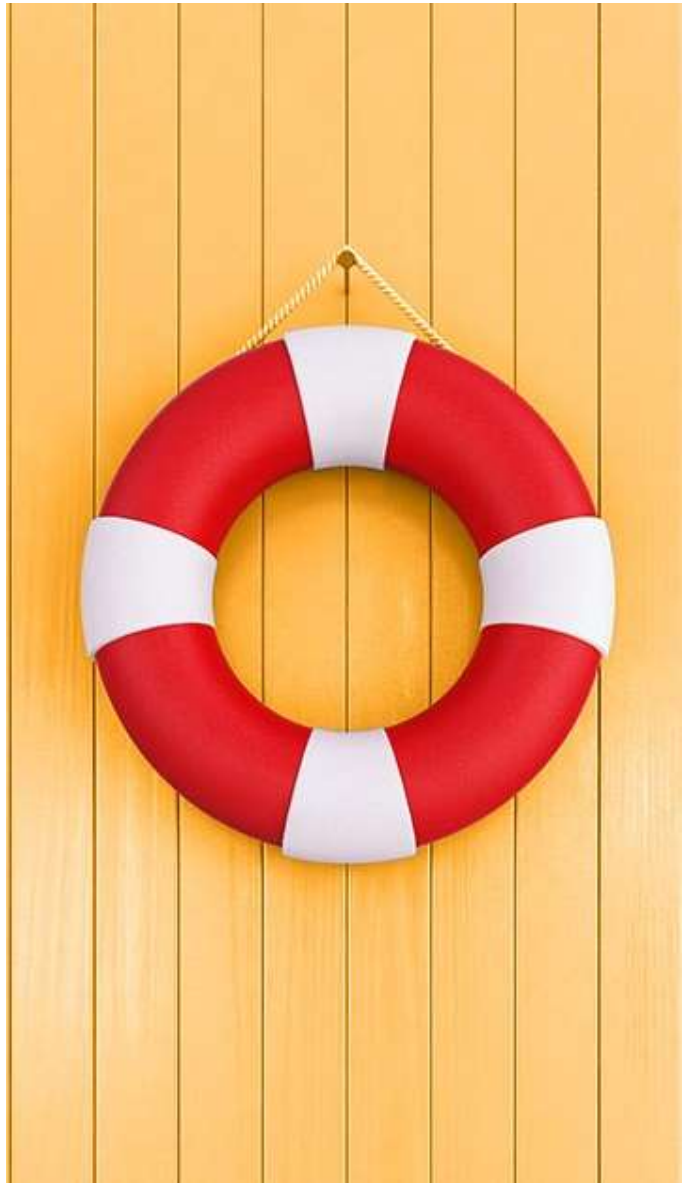
General Assembly: DATR-318

Journal Article Submission Email Engagement



The Problem

SAGE Publications is a publisher of thousands of academic textbooks and over a thousand academic and professional journals. The latter is undergoing a sea change in its business model, from traditional (pay-to-subscribe) to open access (pay-to-submit) meaning a more reliable, institutional-sales based revenue stream must be replaced with an individual-marketing based revenue stream

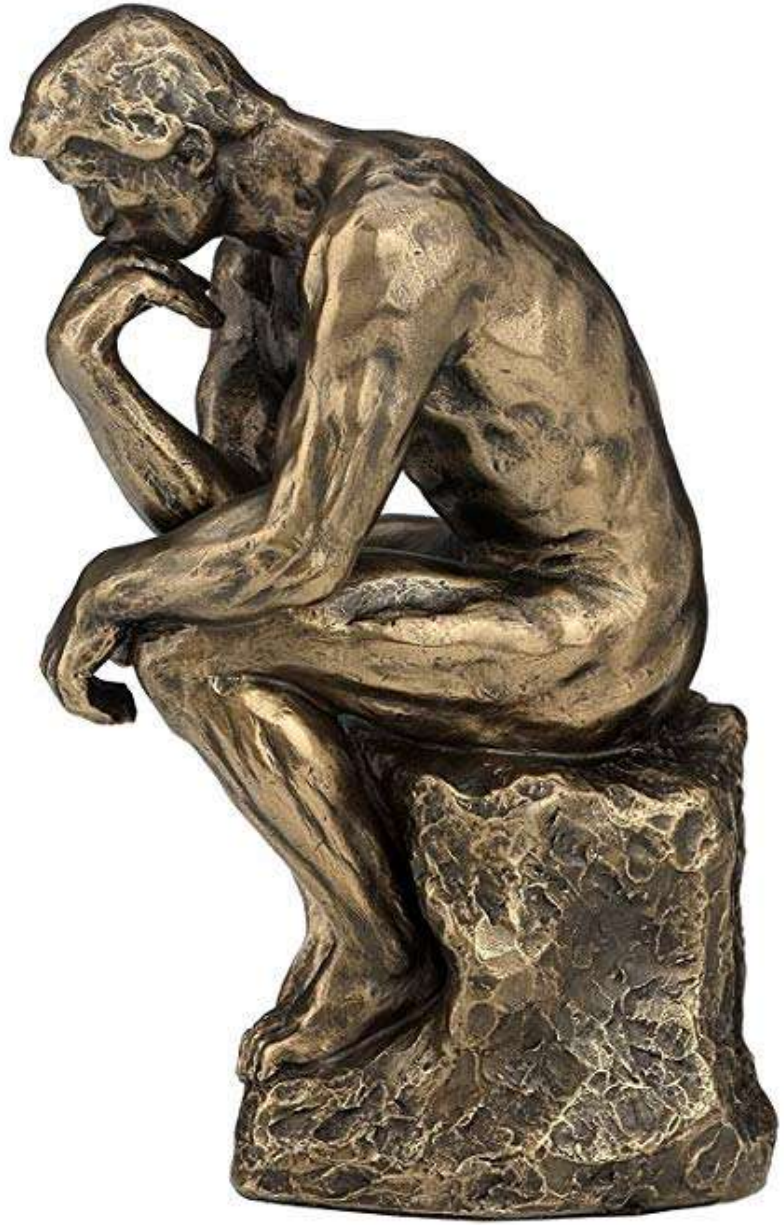
In 2018 the Open Access team sent over 89 million emails to 2.5 million contacts (averaging 35 pp) which is already causing concerns of over-mailing, so the answer cannot simply be to keep increasing volume, rather in a more intelligent approach to audience selection and content targeting. As a result I have built a series of models to predict email engagement based on features that currently are being used for selection, and those that are not but could be included

Hypothesis

Prior simple analysis indicated that lifetime and recent responses to a given product was a potentially valuable factor in future engagement, however these metrics are not easily obtainable for list selection. The development spend required to do so requires sound evidence to make a business case for this enhancement

Using email data for one of the leading Open Access journals for Q1 of 2019, along with features that are primarily used for list selection (submission activity, recorded customer interests etc.) AND features of previous response behaviours, we can train an algorithm to accurately predict the fact of engagement

From this we can hope to identify which features are key to email engagement, and thus determine whether there is evidence for the proposed enhancement



Approach & Data

In order to effectively train my model, I scaled up the initial sample dataset used from 200k to the full data of just under 1 million rows, changed feature flags into feature counts, and from these derived response rates

The data is composed divided into train and test sets, the train containing 699k rows, test containing 233k rows, containing 20 feature variable columns and the target variable of present engagement. This data was manually aggregated from existing and former email campaign sent and response data

	is_academic_email	is_china_recipient	prior_received	recent_received	prior_responses	recent_responses	prior_submissions_subject	prior_submissions_discipline	recent_submissions_subject	recent_submissions_discipline	...
0	0	0	189	16	9	0	0	0	0	0	...
1	1	1	6	7	0	0	0	0	0	0	...
2	1	1	2	3	0	0	0	2	0	2	...
3	1	1	155	23	15	0	0	3	0	0	...
4	1	1	75	36	69	30	0	3	0	0	...

5 rows × 21 columns

Scoring the models

As the problem is one of classification I have used r2 accuracy scores to measure model performance against the null model performance of 78% for the following models:

Logistic Regression:

Train accuracy: 87.88%
Test Accuracy: 87.88%

Decision Tree Classifier:

Train accuracy: 88.64%
Test accuracy: 88.49%

Random Forest Classifier:

Train accuracy: 88.99%
Test accuracy: 88.40%

Conclusions

- While the Random Forest Classifier produced the best predictive result, given the business needs interpretability takes precedence, meaning that the Decision Tree Classifier model is the most appropriate to employ (as it allows a visualization of the ‘flowchart’ of the model)

- Whether looking at feature correlation in Logistic Regression (which factors most closely relate to a recipient engaging) or the feature importances of Decision Tree and Random Forest models (which factors are most valuable in predicting engagement) previous product responses and derived response rates are consistently the highest ranked features related to a given engagement

- Drilling down into the key features correlating to engagement reveals measures that may be practically applied in campaign selection (e.g 50% of recipients who have > 12 recent responses are predicted to engage)

Next steps

- Caveat: In the late stages of modelling I discovered some odd response rates showing (> 100%) meaning that some errors have crept into the data. These should be investigated and the model retrained before drawing final conclusions

- SAGE publishes Open Access journals over many different disciplines, which may have substantially different audiences (Humanities, Medicine, Engineering) and behaviours. I intend to build discipline specific models which in combination will serve as the business case for proposed data enhancements

- Another aspect surfaced by these models is that those contacts who have not previously engaged with a product engage at a significantly reduced rate than those that have. However if we take this as a determining measure then no new names would ever be contacted. As a result I intend to build wider-scale models on this audience group in order to determine the features that correlate with their engagement

- Likewise, the Asia Pacific region is a valuable market for journal article submission, accounts for ?% of this dataset, and engages at a lower rate than the RoW. Region specific modelling may surface different significant features for this audience group (initial investigation indicates differing thresholds are present and may be specifically employed)