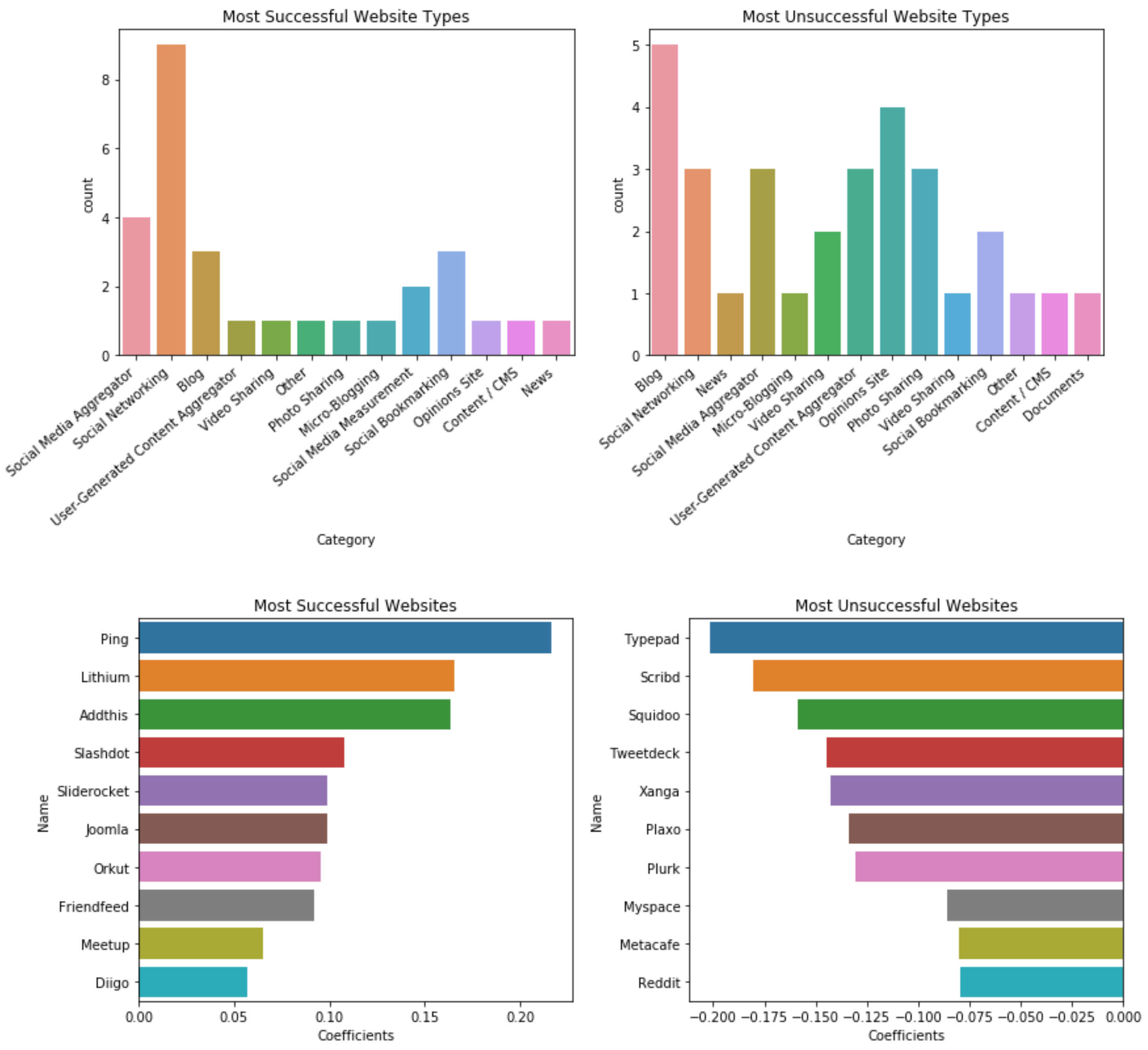
Ad Clicks Technical Challenge

Background

XYZ Limited have carried out an experiment, advertising their service for the last 10 weeks across various websites. The CMO wants a recommendation as to a longer term, more focused, digital investment strategy.

Objectives and Model Background

Given the dataset attained by XYZ Limited, it is critical to investigate which websites are contributing the most towards the successful clicks generated. To do this, I created a model which aims to predict whether a user is likely to click on an ad or not given their website viewing data. The model used was a logistic regression model hich essentially calculates the probability that a user will click on an ad or not. The higher the probability is for a user, the more likely the user is to click on an ad. In developing the model, weights are assigned to websites depending on whether those websites contribute positive/negatively towards ad clicks or not. Websites with higher weights contribute more towards clicks than websites with lower weights. The figure below shows the frequency of the types of websites that were both successful and unsuccessful in generating ad clicks and the top/worst performing websites over the 10 weeks of experiments.



Most Successful Website Types

Most Unsuccessful Website Types

Most Successful Websites

Most Unsuccessful Websites

Recommendations:

Based on these results, the recommendations for the CMO are as follows based on the assumption that the same ads were used across the various websites hence, ad features had no influence on the results.

1. Focus on Social Media related websites. 7/10 of the top performers were related to Social Media in some way e.g. social networking, social aggregators etc. However, focus on the top performers e.g. Ping, Lithium, Addthis as some of the worst performing websites include Social Media websites e.g. Scribd.
2. 6/9 blogs, 3/4 photo sharing, and 4/7 Opinions Sites demonstrated a negative/no contribution towards ad clicks. Perhaps carry out controlled experiments by removing these websites to see confirm whether or not these websites perform poorly especially for the worst performers e.g. reddit, Xanga.
3. There are only a small number of some categories of websites e.g. Documents, Content. It may be worth investigating further i.e. try out other similar websites and see what effect that has on the model and the number of ad clicks.
4. Over 99% of impressions are from Newsvine. The model demonstrated that Newsvine has a very insignificant impact on ad clicks. Reduce investment in Newsvine and observe the effect on ad clicks. The client has the potential to save up to £222,600,000 with the assumptions that the cost of 1000 impressions is £2.31 and there were over 96,300,000,000 impressions on the website.

How confident can we be with these findings?

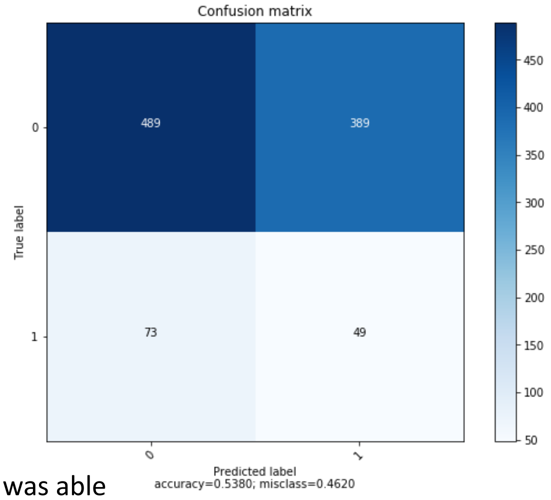
Confidence in the weightings

The model was trained 10 times using a technique called Cross Validation which helps in estimating how well our model will perform in practice. Over the 10 experiments, using statistical measures, it was shown that

on average the percentage error in the average coefficient weightings obtained from the 10 experiments was 0.8%. This means that the weightings of the websites did not change significantly throughout all the experiments carried out. The biggest error was observed for the mean weightings for "Addthis" website at 2.7%.

Confidence in the Model

The figure below shows the number of correctly classified/misclassified predictions in the model. The left axis shows the actual outcome (i.e. ad click or not) and the bottom axis shows the predicted outcomes for example, the bottom right corner shows that 49 samples were correctly identified as an ad click (with 1 denoting ad click). Overall, the model has demonstrated the capability to recall 40% of the ad clicks successfully. Given the error margins calculated from the model, at best the model may be able to recall 67% of the predictions correctly and at worst 31%. The downside to the model is that, it is not extremely precise as the model overestimates the number of ad views attained from the given dataset. However, the model was able to correctly identify most samples which did not lead to an ad view. Thus, the model can shed some light onto how significantly websites are in terms of ad clicks.


Confusion matrix
True label
Predicted label
accuracy=0.5380; misclass=0.4620