**XYZ Limited Recommendations**

# **Modelling**

From the user activity data, I created a robust model to predict customer click-through based on individual websites. I determined that my models based on website categories did not give accurate enough predictions—I lay the evidence of this out below.

Each model was tested to determine its accuracy in the following way:

1. The total 10,000 rows of customer activity data were split into two randomly assigned groups—Group 1 (9,000 rows) was used to train each model, and Group 2 (1,000) to test its accuracy using the available data. These two groups were used for each model.
2. Each model’s predictions were compared against the data, to determine how well they worked.
3. The results of testing the models was then used to establish which model gave the best predictions.

Once I determined which model was the most accurate, it was retrained using all of the available 10,000 rows of data, before being used to make predictions for the additional user data. Its predictions form the basis of my recommendations below.

## **Accuracy of Models**

In total, seven models were tested, using three different techniques. Each technique was used both for category-based and website-based predictions.

As a demonstration of the superiority of website- over category-based predictive models, I have included some figures which show accuracy levels. Below are some ‘Confusion Matrices’—these show how many false positives and negatives a model gave for the test data, shown in the top-right and bottom-left boxes, respectively. A perfect model would have 0 values for these boxes. The titles of each figure relate only to the data-science technique used, and are not essential to understand.

The Confusion Matrices for the two most accurate **category-based** models are shown below. Note that they generate 93 and 106 false predictions per 1,000 data points, and are hence only 90.7% and 89.4% accurate.

Chart

Description automatically generatedChart

Description automatically generated

Compare these with the Confusion Matrices for the two most accurate **website-based** models (including that for the final model, shown on the right). These generate 121 and 47 false predictions per 1,000 data points, and are hence 87.9% and 95.3% accurate.

Chart

Description automatically generatedChart

Description automatically generated

## **Final Model**

The final model used employed a data-science technique called Linear Support Vector Machines (SVM). This works to transform the data into a higher-dimensional space (see figure below), where it can be more easily split up. The graph below shows a transformation from 2D to 3D, whereas my model transformed from around 80 dimensions to over 100.

Diagram

Description automatically generated

In my particular model, the red dots would correspond to a click-through, and the green to a lack of one.

It should be noted that, as with all other models tested, this model assumes that the training data is a representative sample of the global distribution.

# **Recommendations**

A primary recommendation is to avoid basing any decisions on website categories. Allocating resources on a new website on the basis of successful advertising on a similar type of website, for example, is unlikely to be effective. This is evidenced by the fact that none of my category-based models could successfully predict click-throughs.

My website-based recommendations all derive from the most accurate model employed (SVM), which had a successful prediction rate of 95.3%. The graph below lists the 20 most effective websites for advertising, around a quarter of the total websites. The value ‘Model Coefficient’ is a measure of the importance of each website in determining whether a customer clicks-through on an advert; a higher value indicates a higher importance.

From the graph, you can see that the two most important websites are Thisnext and Customerlobby—resources should be focused on these two, with additional resources across the other websites listed. The high accuracy level of the model means that there is a very strong correlation between a customer seeing adverts on these websites, and the probability of them clicking on the advert.

Chart, bar chart

Description automatically generated