Relax Takehome Challenge on User Engagement

In this Challenge, an *adopted user* is defined as a user who has logged in on three separate days in at least one seven-day period. For computational simplicity, this definition was slightly modified here to be a user who had logged in three times in any one of the 52 distinct weeks of the year, over the approximate two-year period under consideration. Of the 12,000 users, only 8,823 seemed to have the requisite timestamp data, and of these, 1,432 met the adopted user requirement (a rate of 16.2S%).

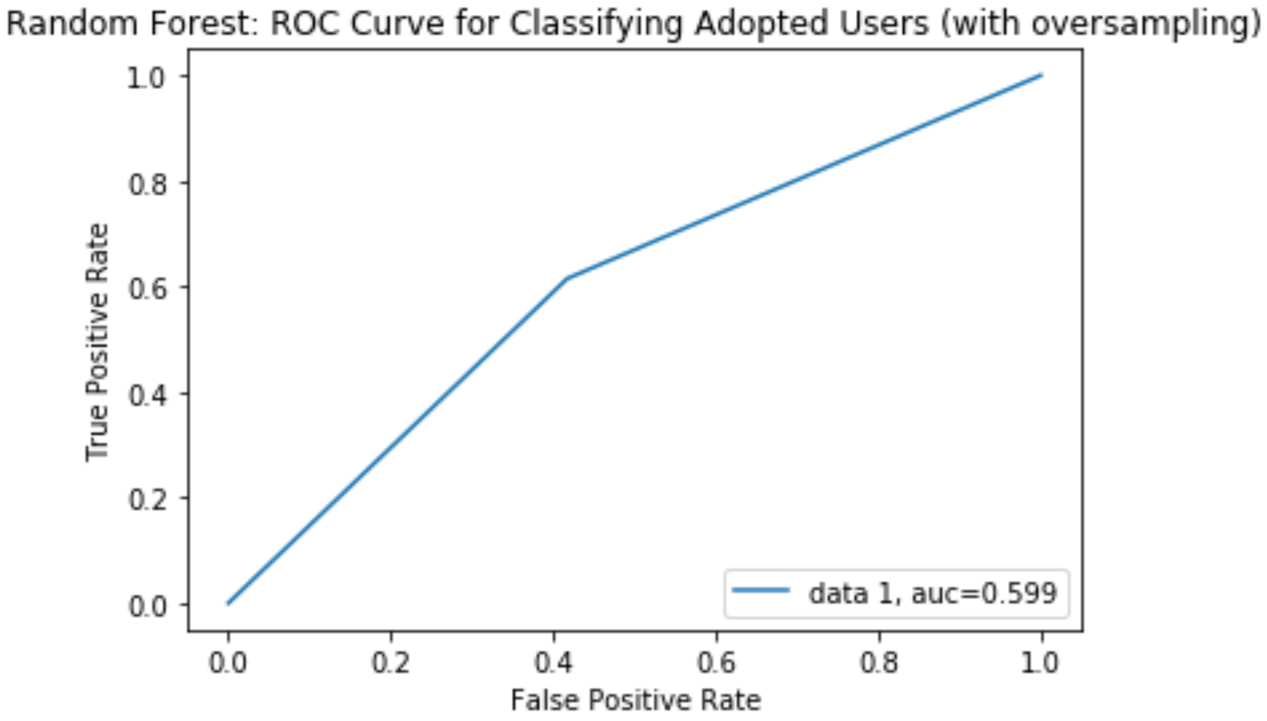
The variables considered were all categorical in nature, being respectively: (a) the five different creation sources, (b) whether the User had opted into receiving marketing emails, (c) whether the User was on the regular marketing email drip, (d) the user's organization ID (with 413 unique values), and (e) the ID of the user who had invited that member to join (or whether no such invitation had been made) – with 2229 unique values. While the variables "*creation_time*" and "*last_session_creation_time*" would likely have been statistically significant in predicting whether a user was *adopted*, these two variables were omitted as being essentially tautological, since a *later last* time would likely be predicted by being 'adopted'. The focus here was therefore on an examination of the various marketing channels, and whether some were particularly important.

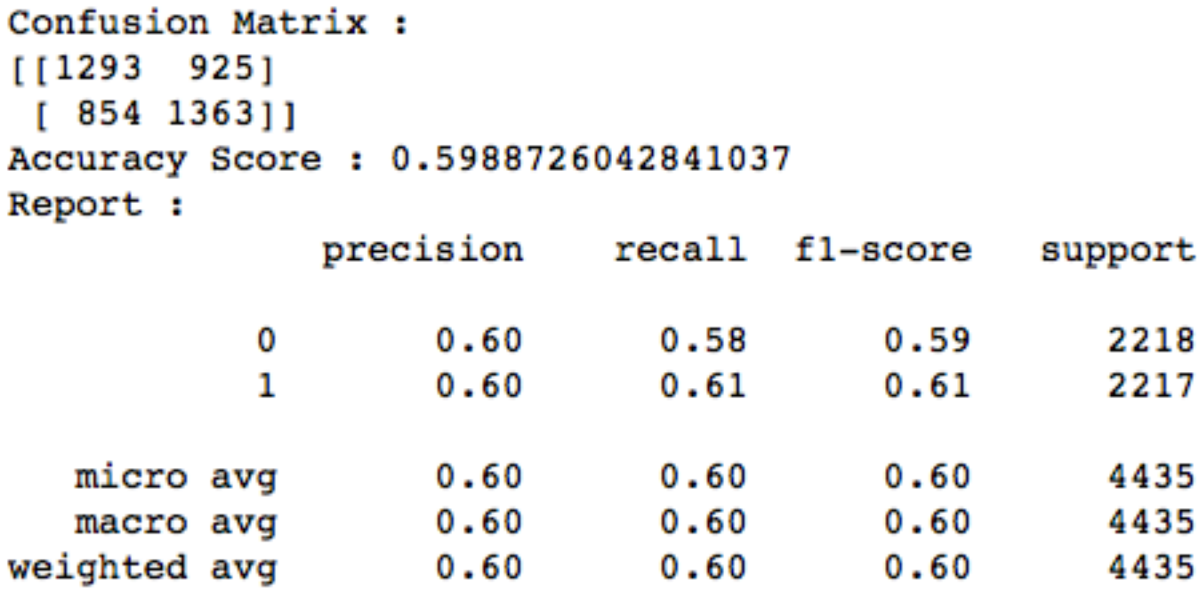
Since this dataset is clearly imbalanced, a Decision Tree classifier ('DTC') was first used, as it is less sensitive to such imbalance. While the "Test" Accuracy of the DTC was superficially impressive at 84.3%, one may note that simply predicting the majority class could achieve an accuracy of 83.8%. Accordingly, the *Area Under the* [ROC] *Curve* was chosen as a better metric for this imbalanced data set, but a value of only 0.499 was recorded, which is almost exactly the 0.5 level that would be achieved by just making random guesses. Accordingly, this DTC model was not considered to be very predictive. Other models (including a *Random Forest* classifier, which is commonly less susceptible to overfitting), were also evaluated, but not found to be much better (with an AUC of 0.500). In order to address potential problems arising from sparse variables, the last two variables cited above (with many sparse values) were dropped. However, the AUC did not improve.

In order to attempt to remediate the class imbalance problem, oversampling was done using scikit-learn's *resample* function. A Random Forest classifier did then achieve an AUC of 0.60, which was the best recorded of numerous classification attempts, but most analysts would not consider such a model especially valuable. Most fundamentally, with many Classification models it is at least possible to classify a small fraction of subjects with a high degree of confidence (so a high *True Positive* rate, and a low *False Positive* rate): however, that does not seem to be the case here, as the gradient of the ROC curve remains uniformly shallow.

Summary: although errors and oversights by this analyst are certainly possible, none of this initial preliminary analysis indicated that the provided variables were especially constructive in predicting whether a User would become "adopted". It may therefore be worth investing in gathering a richer dataset, including variables on the age, gender, income level (using a home address zip code as a proxy) of users. While using immediately available variables may be the fastest and cheapest route, if these variables are not predictive then investment in additional information gathering may be justified. There is also the possibility that what is of paramount importance is simply the User's experience of the Relax Product, and how the user happens to be introduced to the product is not especially important. This would actually be a valuable conclusion in itself, as the business implication would be to not bother investing in promoting particular marketing channels, and rather to focus on improving the product itself.

While a number of Jupyter notebooks were used in the analysis, only the last one (with oversampling to try to remediate the class imbalance) is provided (at: https://github.com/Methanogen/2019-Springboard-Projects/blob/master/relax_takehome6F%20w%20upsample.ipynb).




```
Confusion Matrix :
[[1293  925]
 [ 854 1363]]
Accuracy Score : 0.5988726042841037
Report :
              precision    recall  f1-score   support

           0       0.60      0.58      0.59      2218
           1       0.60      0.61      0.61      2217

   micro avg       0.60      0.60      0.60      4435
   macro avg       0.60      0.60      0.60      4435
weighted avg       0.60      0.60      0.60      4435
```

Supplemental Thought:

Is this the right question? As one looks at the data in the context of Exploratory data analysis, something interesting emerges. The company has a large tail of high-frequency users, who may well contribute the bulk of the company's aggregate revenues. For example, 90% of the aggregate visits (and an even higher percentage of repeat visits) came from just 11% of users (numbering 1,000), who visited at least 36 times. This may be a more economically relevant group to look at, rather than users who have merely once visited three times in a single week over the two-year period. Further work might address this. However, if the reality is that the product itself is of paramount importance, then this variable would also not be helpful.