# Relax Challenge Report

**By Scott W. Lew**

## Introduction

Relax is a website/service that compiles data on their user logins. The company is interested in the characteristics of their adopted users that lead to a long-term relationship with their company.

An "adopted user" is defined as a user who has logged into the product on three separate days in at least one seven day period. In this project, the goal is to identify which factors predict future user adoption.

## Exploratory Data Analysis

The user data is provided in two csv files: one file has login times for each customer and the second file has customer data with various features such as the customer name and the organization they belong to. Data from both files were loaded into Pandas dataframes and eventually merged into one dataframe.

Organization 0 with 228 users had the most users of all the organizations in the data. And, organizations 4,1,and 7, which had the most adopted users of all the organizations, had 17,16 and 16 adopted users respectively.

Relax users who received an organization invite to join Relax were more likely to become adopters in comparison to other creation source methods.

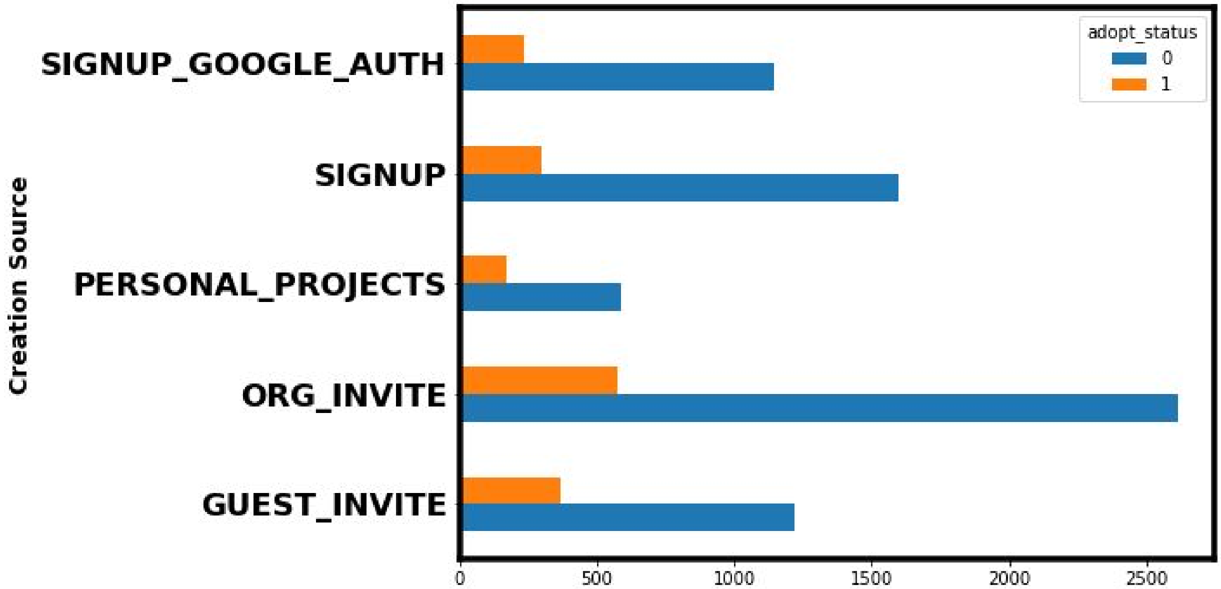


**Figure 1 Creation source of adopted users vs non-adopters**

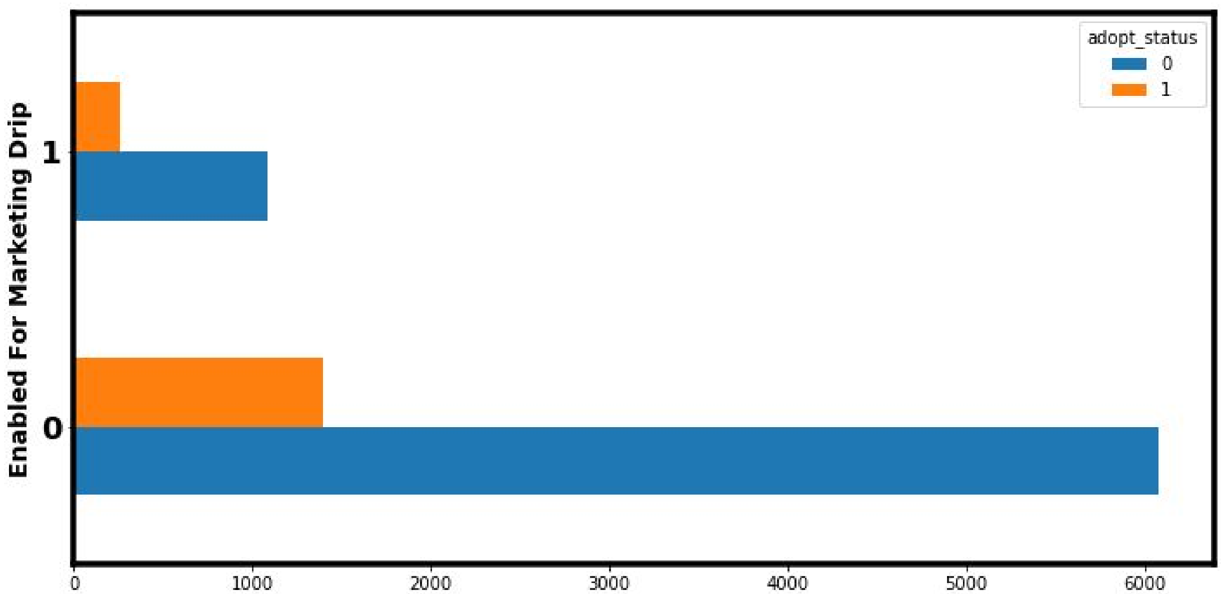


**Figure 2 Enabled for Marketing Drip status of adopted users vs non-adopters**

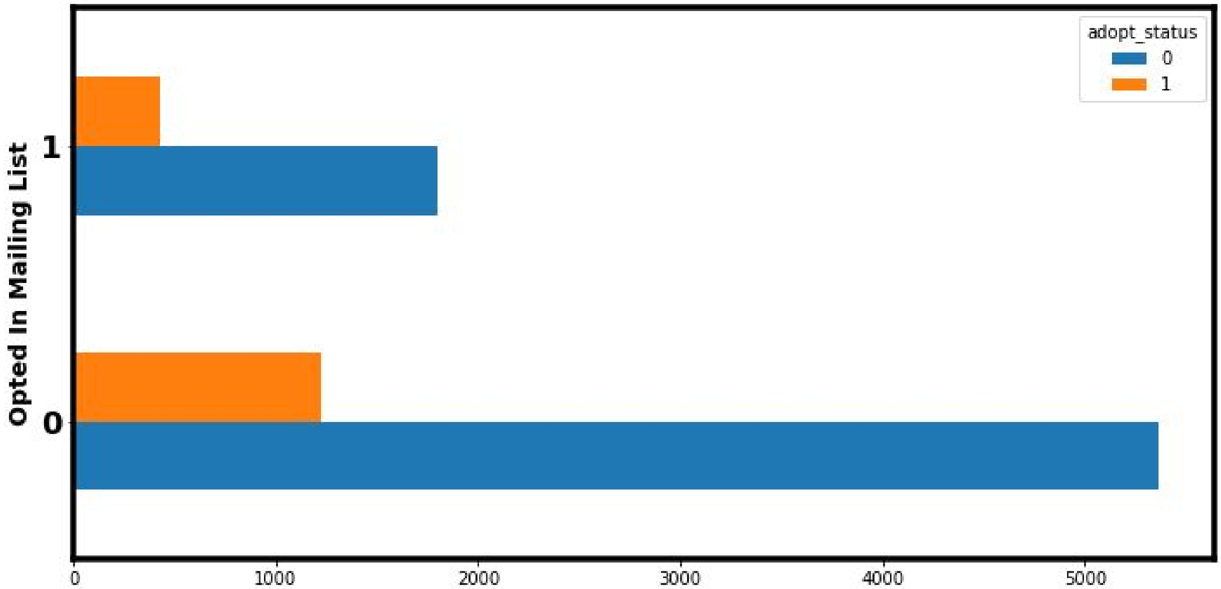


**Figure 3 Mailing list option of adopted users vs non-adopters**

Relax users who are *not* enabled for marketing drip are *more* likely to become adopters. And, Relax users who opted *not* to receive marketing e-mails are *more* likely to become adopted users than those who receive the e-mails. It appears that Relax's e-mail marketing methods are not effective in motivating customers to become adopters.

## Methodology

A user-defined function called **simple_check_2** was created to be used with the Pandas data frame of Relax users in order to determine which user was assigned an adopted user status. As a reminder, an adopted user is one who has logged into the Relax website at least 3 times in any given 7 day period. The function **simple_check_2** performed a group by user_id, and checked login datetimes for each user and calculated how many times he or she logged in 7 day intervals. Those users who logged in 3 or more times in a 7 day period were given adopted user status while those who did not were not considered adopters. After creating labels using the simple_check_2 function, the two data frames of customer data were merged on the common column, user_id.

From the data,  there are 4.5 times more non-adopters than adopters for the Relax website.

## Classification

In order to determine which features are significant in predicting whether a user would adopt the Relax site, an XGBoost classifier was employed for supervised classification. Initially, Logistic Regression, Random Forest and Decistion Tree classifiers were used, but the XGBoost method gave the best results in predicting adopters. As shown in Figure 4 below, the most important features as determined by XGBoost were in descending order of importance: last_session_creation_time, the creation_source, org_id, and enabled_for_marketing_drip.

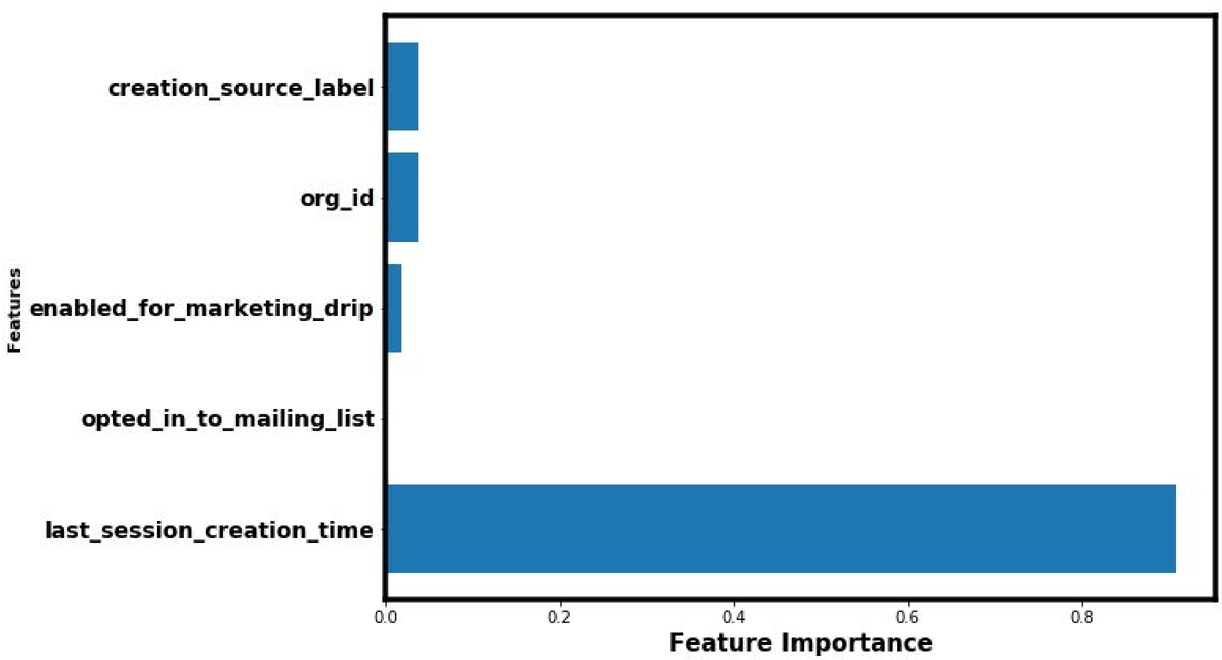


**Figure 4 Feature importance of XGBoost Model**

## Classification Metrics

This XGBoost model has an AUC (Area Under the Curve) score for the ROC (receiver operating characteristic curve) of 0.78.

**TABLE 1: Classification Metrics for XGBoost Model**

| Adopt Status | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0.96 | 0.91 | 0.93 | 1495 |
| 1 | 0.61 | 0.79 | 0.69 | 270 |

The F1 scores produced by the XGBoost model for adopters and non-adopters are 0.69 and 0.93 respectively.

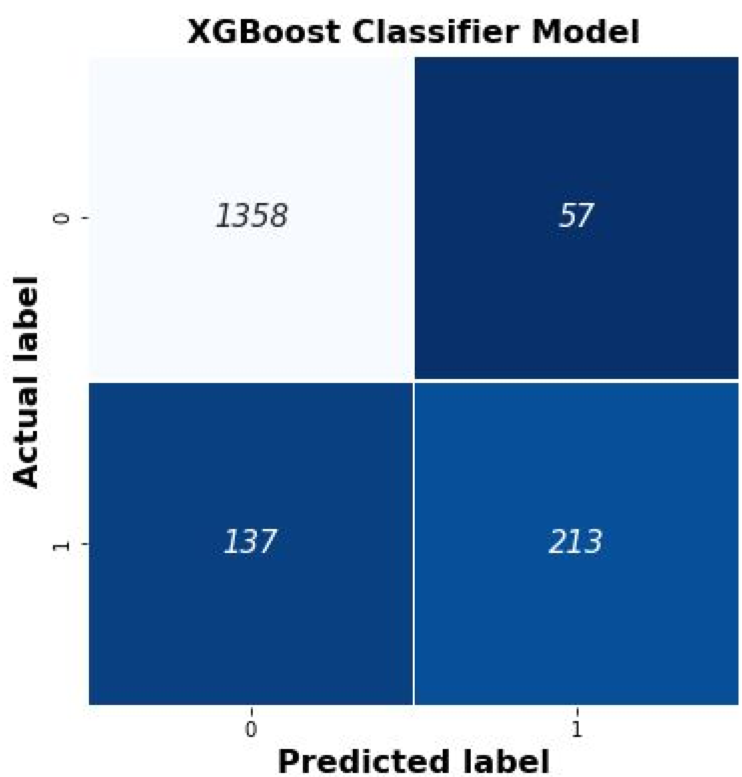


**Figure 5 Confusion Matrix of XGBoost Model**

As shown in the Confusion Matrix above (Figure 5), 1358 customers were correctly classified as non-adopters and 213 customers were correctly classified as adopters. But, the model incorrectly classified 57 customers as adopters and incorrectly classified 137 customers as non-adopters.

The last_session_creation_time, the unix timestamp of last login, is the most important feature in predicting whether a customer will become an adopter user. In addition, the organization that the user belonged to is also an important predictor of adopter status: certain organizations favor using Relax than others.

## Summary

The goal of this project is to determine what data is useful in predicting which customers of Relax will become long-term users, adopters. To address this goal, Machine Learning models were utilized, and the XGBoost method gave the best results. The last_session_creation_time is the most useful factor in predicting an adopted user. It was also determined that certain organizations favored Relax more than others, and the members of those organizations are more likely to become adopters. Based on the exploratory data analysis above, the e-mail marketing used in the past is not effective in creating adopters. Therefore, a new approach to e-mail marketing may be needed to increase customer adoption.