Relax Data Science Challenge

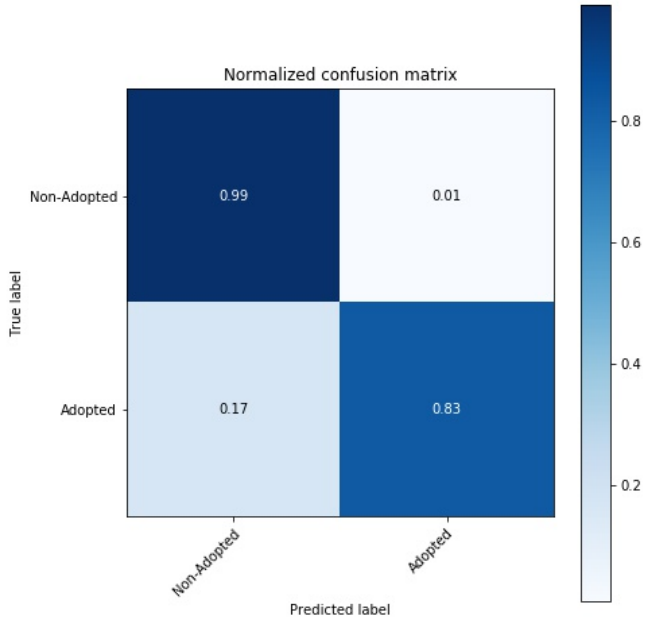
Springboard Data Science
August 2019
Ishmail Grady

The goal of this exercise is to determine what factors predict future user adoption using data on 12,000 users who signed up for the product in the last two years. To accomplish this goal a logistic regression model was developed to predict if a user is an adopted user or not. The following features were considered:

- Account Age – the time in days since to the account was first created (numerical)
- Days Since Last Session – the time in days since the user last used the product (numerical)
- Creation Source – the method used to create the account (categorical)
- Opted in to Mailing List - whether the user has opted into receiving marketing emails (Boolean)
- Enabled for Marketing Drip - whether the user is on the regular marketing email drip

Numerical features were scaled to have a mean of 0 and standard deviation of 1. Categorical features were encoded using one hot encoding (last feature in each category was dropped). The dataset contained 3,177 rows with null values; these records were discarded. The model was trained using 80% of the remaining 8,823 records and 20% for testing.

The final model achieved accuracy of 96% on the test data. As shown in, the normalized confusion matrix for the testing results below, the true positive rate for adopted users is 83% and the false negative rate for adopted users is 1%. These metrics indicate that the model performs the task of predicting adopted users and give confidence that it will perform well on future data.



Given that the model performs well, we can use it to determine what factors predict future user adoption. The top 2 factors are the days since Days Since Last Session and Account Age. The Days Since Last Session feature has a negative coefficient while Account Age has a positive coefficient. This indicates that older accounts with shorter periods of inactivity are more likely to be adopted users.