The main target used here is the “adopted user” criteria that checks to see if a user will log into the product 3 days within any 7 days period. First, the login information was used to create a new data frame which contains the user id and a feature called “adopt” which takes on the values 0 or 1 depending on if the user is defined to be not adopted or adopted respectively. Before merging this new data frame with the user information data frame new features were created using data from the user data set. The creation source was given a number for each unique source (5 unique sources) and a number designated for each unique email domain (1184 unique domains). Also, the account creation time was broken up into creation year and creation month to add additional features. Next the adopted data frame was merged with the user data frame with new features. The resulting data frame has a total of 8823 observations in it which is the same as the number of users in the data. The only column with null values was the “invited_by_user_id” however it is determined that this column should not have bearing on the prediction algorithm and hence the null values are ignored. The final features used for model was:

```
['last_session_creation_time', 'opted_in_to_mailing_list',
'enabled_for_marketing_drip', 'org_id', 'email_type',
'creation_source_num', 'creation_year', 'creation_month', 'user_id']
```

The final target, adopted user value, is binary and therefore this is a simple binary classification problem. Because of the number of features being used and the possibility for non-linearities in target response to features a decision tree method is decided for use in this problem. Ultimately, a the RandomForestClassifier() and the boosting decision tree algorithm XGBoost() are used to solve this problem. A grid search with 5-fold cross validation was used to tune modeling parameters. The final accuracies for classification on the test set and the resulting ROC curves (with AUC calculations) was:

```
Random Forest Accuracy 0.942

XGBoost Accuracy 0.962
```

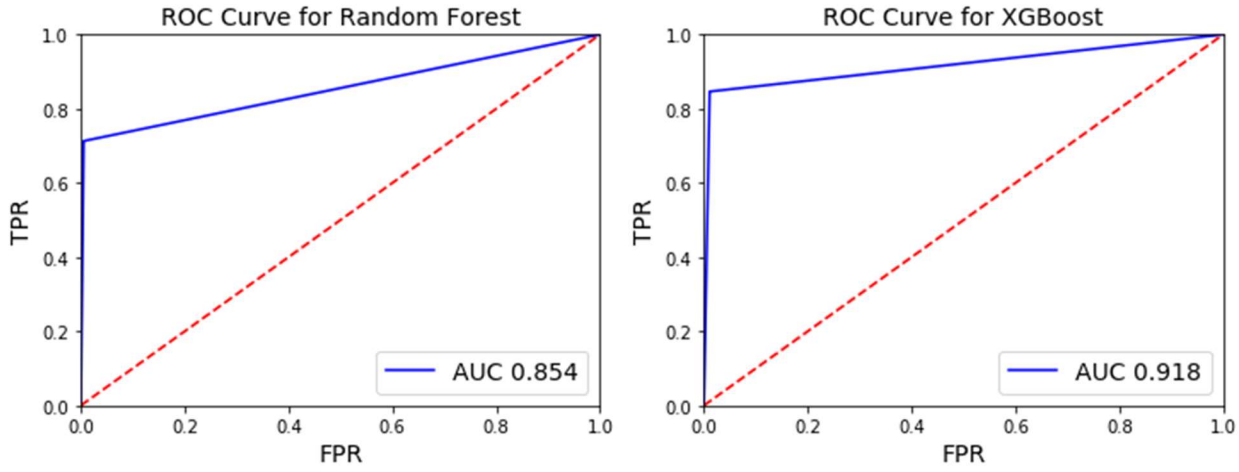

Figure 1 ROC curve for the rando forest and XGBoost models.

This indicates that the prediction of whether a user is adopted or not is accurate using the features and models in the report. The next step is to look the effects of different features on the prediction.

Here, the feature importance for the random forest and boosting algorithm are shown (Figure 2). Notice, in both models the last session creation time is a very important factor. Hence, the more recently someone has signed in the more likely they are to be an “adopted user”. Also, there is a strong correlation between the account creation year and month with if the user will become adopted. These trends can be seen in the bar charts that show different feature quantities with the adoption fraction within that feature grouping. A few trends and recommendation can be made from this information.

- Surprisingly, the two marketing methods used (marketing drip and email list) did not have a significant effect on the adopted user fraction. Putting more effort into these methods would not be advised.
- The last login time is important which is naturally true because more frequent logins typically would translate to more recent logins (see on page 4 where last login time is removed as feature). There may be little improvement that could change this.
- It is noted that the “adopt user” fraction is dropping with more recent years. Perhaps this is due to lack of early advertising of product more recently. Perhaps put a promotional offer for users early to see if it boosts logins. It could also be due to increased competition with other products in recent years which would mean more effort needs to be put into retaining users. Exploring more in depth to why adopted user fraction dropped from 2012 to 2014 would give better ideas about recommendations.
- For some reason, late spring (April and May) has a sharp drop off in adopted user fraction. Research and/or surveys should be used to understand this trend to better make recommendations.

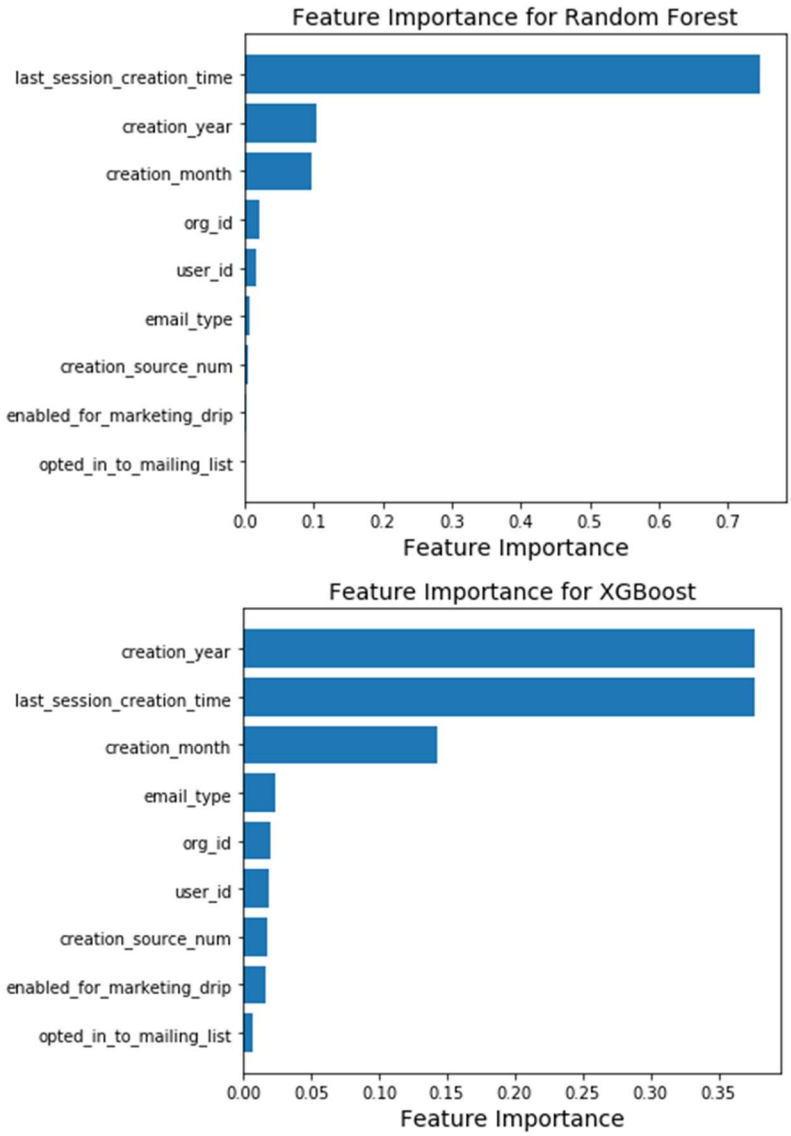

Figure 2 Feature importance in the random forest (top) and XGBoost (bottom) algorithms.

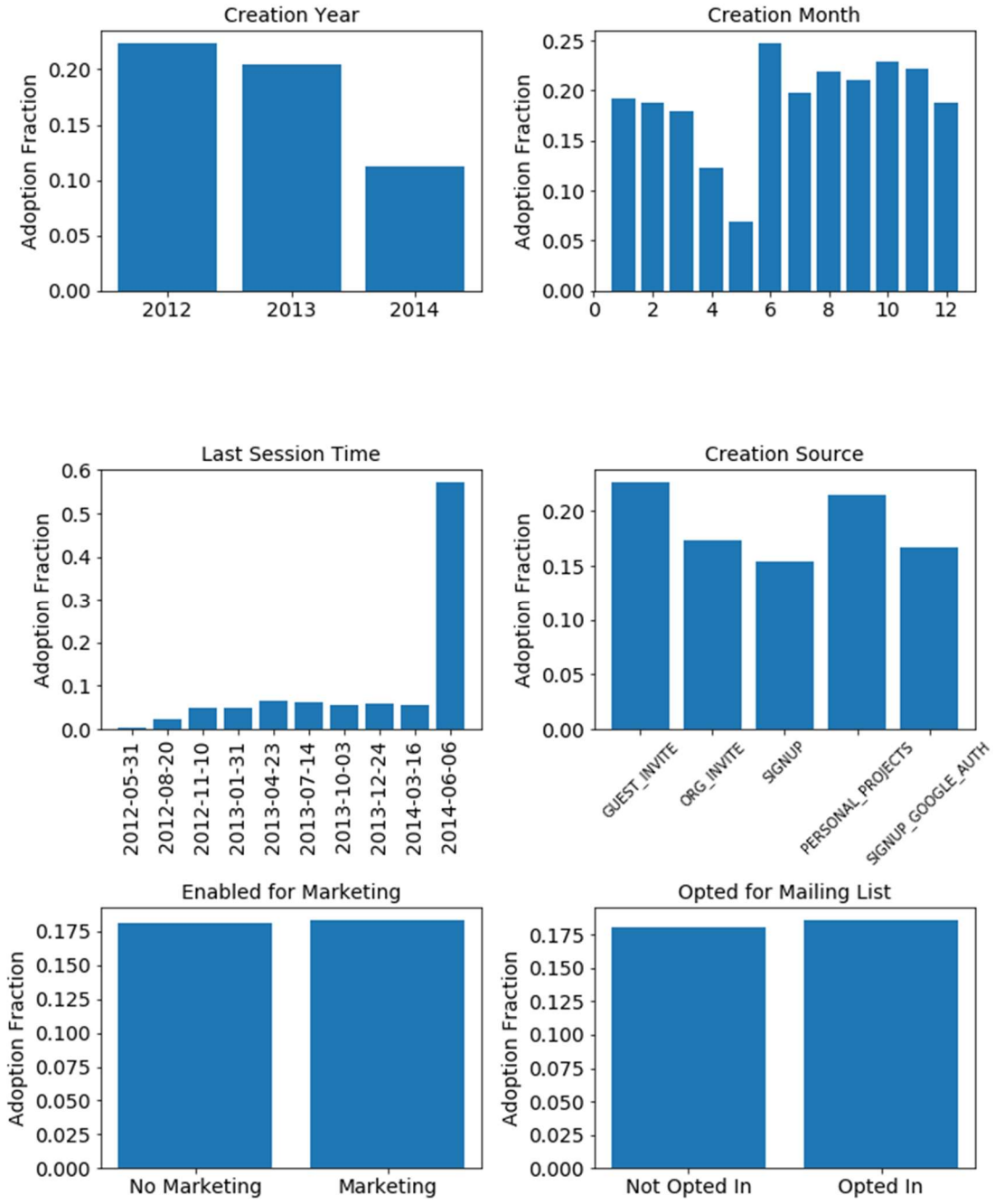

Figure 3 Bar charts for different features displaying how feature values change the fraction of users that are "adopted users."

One thing that is curious is the significance of the last session creation time. This factor may be too closely related to the target and therefore could block other significant factors in the prediction of the adopted users. Here, we simply show how the model changes if the last session creation time was removed. The accuracy for the model and the feature importance is shown below. Notice, that the accuracy decreases when removing the last login time and the hierarchy of feature importance is partly changed. The account creation year and month both are still highly important features but the origination id increases in the hierarchy of importance. This merely suggests that without the last login time available then the likelihood of user adoption is largely based on account creation date and partly due to the origination they below to.

```
Random Forest Accuracy 0.813

XGBoost Accuracy 0.813
```

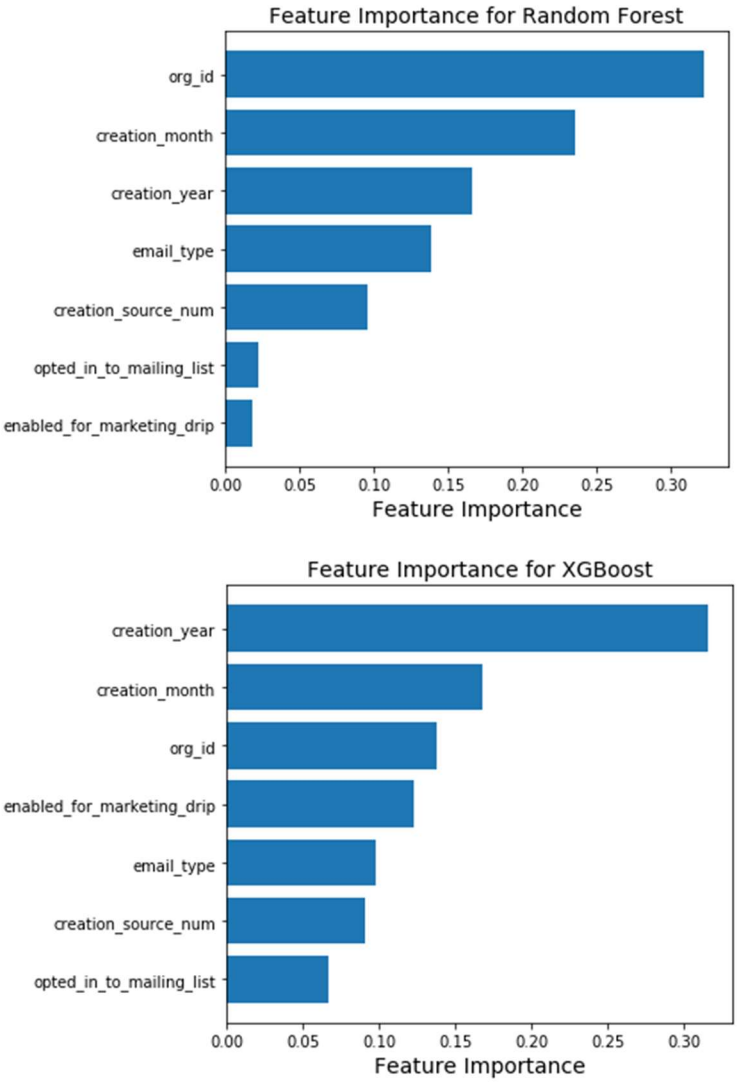

Figure 4 Feature importance for random forest (top) and XGBoost (bottom) after removing the most important feature from original model.