Relax Challenge

First step:

Checking the datasets:

The users table conatins data of 12000 users and how their account was created (source) wheras the engagement table has 207917 entries about how the users visited the site. I quickly checked that the ‘visited’ column has no entry other than 1. I also checked the two datasets for discrepancy:

There are 8823 users who have ever used the application.

There are 12000 signed up for the application.

There is a discrepancy of 3177 users.

Second step:

Defining an "adopted user" as a user who has logged into the product on three separate days in at least one seven­day period,

According to my calculations 16.72% of users can be considered to be adopted.

Third step:

I checked how the creation sources are divided:

creation\_source feature:

ORG\_INVITE 4254

GUEST\_INVITE 2163

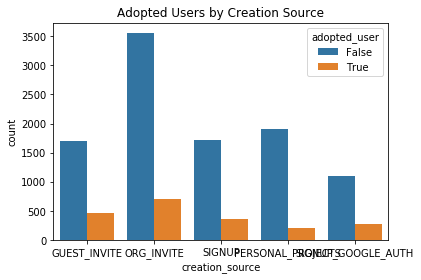
PERSONAL\_PROJECTS 2111

SIGNUP 2087

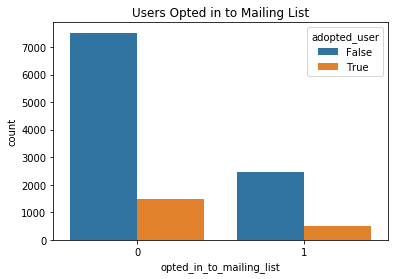
SIGNUP\_GOOGLE\_AUTH 1385

Now three Plots:

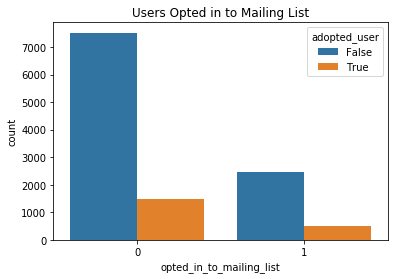
1.



2.



3.



Finding so far:

Only 16.7% of the data contains adopted users

Most common creation source for adopted users is through their own organization's invite

Majority of adopted users did not opt in to the mailing list or were on regular marketing drip

Modeling Approach:

Since we know that the number of times used and the duration between the first and last time defines an adopted user, we would not gain much inference in including those two columns as features in our predictive model. The real question that we have here is: What is driving users to use the product more and for longer periods of time? The key columns that we can develop insights from are the creation source, whether or not they were opted in to the mailing list and if they were part of the regular marketing drip.

However our data does not contain enough adopted user instances to get a good enough insight on whether or not those drive user adoption. With that said, I do not expect us to infer much from the following models that will predict user adoption.

Fourth step:

At this point I started preparing for modelling and figuring out features which will help in modelling. The goal is to determine factors which will predict user adoption.

1. I dropped columns : last\_session\_creation\_time, invited\_by\_user\_id
2. I dropped columns which are more descriptive in nature than predictive like user\_id, name, email, creation\_time,org\_id etc.

Fifth step:

In this step I applied Logistic Regression(LR), Kneighbors Classifier(KN) and Decision Tree Classifier(CART). None of them gave good accuracy score. The number in parethesis is the stadard deviation.

LR: 0.498612 (0.021622)

KN: 0.513786 (0.019566)

CART: 0.559178 (0.015343)

Sixth step:

In this step I applied Synthethic Minority Oversampling Technique (SMOTE)

To take care of class imbalance. In oversampling the minority cass instances are duplicated for overfitting. With SMOTE, I again applied Logistic Regression(LR), Kneighbors Classifier(KN) and Decision Tree Classifier(CART). The scores improved.

LR: 0.600000 (0.489898)

KN: 0.216142 (0.102143)

CART: 0.700488 (0.250743)

Now I applied AdaBoostClassifier(AB), GradientBoostingClassifier(GBM) and RandomForestClassifier(RF) with SMOTE. I got good scores. Gradient Boost gave the best score.

AB: 0.689058 (0.264583)

GBM: 0.697490 (0.254487)

RF: 0.692680 (0.262202)

Seventh Step:

I fit the model with Gradient Boost and plotted the feature importance matrix

Conclusion:

There are a couple takeaways from the data exploration that the company could move forward with:

The source of account creation is the most important feature regarding prediction of an adopted user The majority of adopted users seen here were not on the mailing list or part of the regular marketing drip, which means that these two avenues of company outreach may not be worth continuing The most common source of creation between adopted users was through organization invite, this could be a cue to hone in on outreach within an individual user's organization to develop more user adoption.

1. Since there is imbalance in the adoption field, I used the class\_weight = ‘balanced’ parameter of the classifier.
2. I got a training score of 0.84 and testing score of 0.72
3. With GridSearch, I could get upto 0.75 accuracy score.
4. Here is the orcer of feature importance-
5. feature\_importance

org\_id 0.935802

creation\_source\_PERSONAL\_PROJECTS 0.015795

enabled\_for\_marketing\_drip 0.015736

opted\_in\_to\_mailing\_list 0.014804

creation\_source\_SIGNUP\_GOOGLE\_AUTH 0.006451

creation\_source\_GUEST\_INVITE 0.005424

creation\_source\_SIGNUP 0.003120

creation\_source\_ORG\_INVITE0.002869