Identify factors that predict future user adoption

Candidate Name: Ashlesh Billady Shetty

Introduction:

Every product-based companies spend a lot of money in marketing to reach to its potential user base. Once a potential user is aware of the product he engages with the product, evaluates the product and finally takes a decision to adopt the product. So, to get a good return on investment for the company it is crucial to predict future user adoption and to understand factors that influence user adoption.

In this case 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. User engagement data and the user information data is used to understand factors that have a significant relationship with user adoption and key factors that helps in predicting future user adoption.

Data Preparation:

User engagement table is used to identify the adopted users. Before calculating any metrics from the engagement table, the data for adopted users from the day they became an adopted user is removed, because we are interested in knowing the behavior and factors that are influencing a user before he/she became an adopted user.

Once we have all the key metrics from user engagement and user information table we build a master data with all the key metrics/features at every user level for all users who are present in both the tables

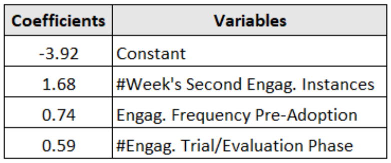
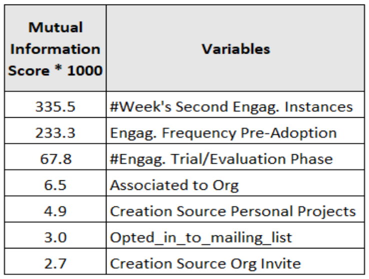
Key Hypotheses:

1. During the first week of engagement (i.e. the trial and evaluation stage of the adoption process) users with higher engagement have significantly different adoption pattern
    - *Metric used:* Number of times user engaged in the first week. (This metric will have a minor flaw only for few users who became adopted users in first week)
2. During the pre-adoption phase (i.e. users who have crossed the trial and evaluation stage of the adoption process) users with high frequency of weekly engagement have significantly different adoption pattern
    - *Metric used:* For every user who have engaged with the product for more than a week, ratio of total engagements to total number of days between the first and the last engagement is calculated. Remaining users are assigned 0.
3. Users who are having high number of second engagements in a week's duration have significantly different adoption pattern
    - *Metric used:* For every user total number of engagements that was the second engagement in the last seven days is calculated
4. Users who show special interest in the product (i.e. users who have opted to mailing list or enabled for marketing drip) have significantly different adoption pattern
5. Users who have affiliation to an organization have significantly different adoption pattern
    - *Metric used:* Users with org ID populated are flagged 1 and rest 0
6. Users from different creation sources have significantly different adoption pattern
7. Users who start their engagement in different time of the year have significantly different adoption rate
    - *Metric used:* month is extracted from creation time variable

Key Hypotheses Tests and Results:

Hypothesis Number	Hypothesis Metrics Short form	Test	Test Statistic	P-Value	Null Hypo. Rejected
1	#Engag. Trial/Evaluation Phase	Chi-square Test of Indep.	1541.262	0.00	**Yes**
2	Engag. Frequency Pre-Adoption	Logistic Regression	0.423	1.22e-22	**Yes**
3	#Week's Second Engag. Instances	Logistic Regression	0.480	5.91e-93	**Yes**
4.A	Opted to mailing list	Chi-square Test of Indep.	0.351	0.55	**No**
4.B	Enabled for marketing	Chi-square Test of Indep.	0.024	0.88	**No**
5	Associated to Org	Chi-square Test of Indep.	27.083	1.94e-07	**Yes**
6	Creation Sources	Chi-square Test of Indep.	40.090	4.14e-08	**Yes**
7	Engagement start month	Chi-square Test of Indep.	155.814	9.64e-28	**Yes**

Predictive modeling:

<table>
<tr><th>Model Built</th><th>Dependent Variable</th><th>Independent Variables Selection</th><th>Model Summary</th><th>Model Evaluation</th></tr>
<tr><td><b>Logistic Regression:</b> For good interpretability</td><td>Adopted user flagged 1 and others 0</td><td>After checking for Multicollinearity, P-Value and intuitive sign on the coefficients three variables remain</td><td>Pseudo R-Square: 0.537
<table><tr><th>Coefficients</th><th>Variables</th></tr><tr><td>-3.92</td><td>Constant</td></tr><tr><td>1.68</td><td>#Week's Second Engag. Instances</td></tr><tr><td>0.74</td><td>Engag. Frequency Pre-Adoption</td></tr><tr><td>0.59</td><td>#Engag. Trial/Evaluation Phase</td></tr></table></td><td>F1 Score:0.91
AUC:0.86
Accuracy: 0.90</td></tr>
<tr><td><b>XGBoost:</b> For good predictive performance</td><td>Adopted user flagged 1 and others 0</td><td>Top 7 variables with highest mutual information score is shortlisted to build the model
<table><tr><th>Mutual Information Score * 1000</th><th>Variables</th></tr><tr><td>335.5</td><td>#Week's Second Engag. Instances</td></tr><tr><td>233.3</td><td>Engag. Frequency Pre-Adoption</td></tr><tr><td>67.8</td><td>#Engag. Trial/Evaluation Phase</td></tr><tr><td>6.5</td><td>Associated to Org</td></tr><tr><td>4.9</td><td>Creation Source Personal Projects</td></tr><tr><td>3.0</td><td>Opted_in_to_mailing_list</td></tr><tr><td>2.7</td><td>Creation Source Org Invite</td></tr></table></td><td>Minimum child weight = 5
Objective = binary: logistic
Learning rate = 0.1
Maximum Depth=3</td><td>F1 Score: 0.95
AUC: 0.96
Accuracy: 0.95</td></tr>
</table>

Conclusion:

From the complete data analysis, hypotheses testing, and model building exercise future it is clear that user adoption prediction can be best done by variables that can capture user engagement characteristics in the trial/evaluation phase and in the pre-adoption phase. Beyond these user engagement characteristics features there are three more features that have significant relation with future user adoption. These three features are creation source, association to an org and month in which user started the engagement.

Next Steps:

There are 3 potential ways to further improve the prediction of future user adoption. First, by obtaining more business context so that potential flaws in the existing model can be addressed. This will also give rise to more hypotheses. Second, with additional data such as product log data etc. will give additional informative variables. Lastly, trying other classification algorithm and generalizability techniques