Walter Piper

Relax Data Science Challenge for Springboard

Notebook at <https://github.com/WalterPiTheScienceGuy/forSpringboard/blob/main/Relax%20Challenge/Relax_Challenge_WalterPiper.ipynb>

Model Metrics at <https://github.com/WalterPiTheScienceGuy/forSpringboard/blob/main/Relax%20Challenge/model%20metrics.xlsx>

**Problem Statement**

Identify which factors predict user adoption. Adoption is defined as logging on three separate days in at least one seven-day period.

**Data Wrangling and Preprocessing**

The dataset from ‘takehome\_users.csv’ contained 4 string/object columns and 5 numeric columns.

* Of the string columns,
  + the “name” and “email” columns were dropped, as these strings were not expected to be predictive.
  + The “creation\_source” was one-hot encoded, and one of the 5 resulting dummy columns was dropped before modeling to avoid collinearity.
  + The “creation\_time” was converted to datetime. Before modeling, the year was extracted from “creation\_time”, while the month, weekday, and hour were dropped because exploratory data analysis did not show predictive validity.
* Of the numeric columns,
  + “object\_id” (equivalent to user\_id) was assigned to the index and removed from the feature columns.
  + The value counts of the column “org\_id” were sorted, and the most common org\_id values that accounted for 50% of the users were all assigned the value “1” in a new binary column “org\_id\_large\_org”. The larger orgs giving the value “1” were orgs that had at least 29 affiliated users.
  + “invited\_by\_user\_id” had a large number of null values, so this column was converted into a binary variable in a new column “was\_invited\_by\_a\_user” with a value of “1” if the value was not null.
  + “last\_session\_creation\_time” was dropped, as this was not expected to be relevant.

For each user, the login dataset (“takehome\_user\_engagement.csv”) was searched with a function that returned the max number of logins within a 7-day period. If this max value was at least 3, the user was considered to have adopted the product. Only 1,597 out of the 12,000 users met this criterion.

**Exploratory Data Analysis**

The mean values for adopters and non-adopters for each feature were compared. The Spearman correlation matrix **(Figure 1)** was also examined to examine feature-to-feature correlations and target-to-feature correlations. While strong relationships were lacking, some relationships appeared. These included a positive correlation between adoption and the GUEST\_INVITE creation method, and a negative correlation between adoption and PERSONAL\_PROJECTS, large orgs, invitations by other users, and the creation year (more recent users were less likely to adopt).

**Modeling**

Because 8 out of 9 features were binary (features shown in **Table 1**), I opted to use decision tree methods instead of parametric methods. These methods also allow easy visualization of feature importances. The tested models included RandomForestClassifier, ExtraTreesClassifier, and GradientBoostingClassifier. These were ranked according to the cross-validation ROC\_AUC score (Receiving Operating Characteristic – Area Under the Curve). Grid search was used to find optimal parameters. There was a large class imbalance that confounded early modeling attempts, so I resampled users with replacement to obtain 10,000 adopting users and 10,000 non-adopting users.

The best model **(Table2)** was an ExtraTreesClassifier with 25 estimators and bootstrap sampling of 85%. However, every model reached approximately 60% accuracy on each of the two classes. This indicates that the available data is not strongly predictive of adoption, even if the 60% accuracy is slightly better than chance (50%).

**Recommendations**

Creation year is the most important feature **(Figure 2).** Users with a longer history are more likely to have adopted the product. Next most important is whether users signed up for PERSONAL\_PROJECTS or if the org they come from has many other users. Both these conditions are associated with less likelihood to have adopted the project. Finally, enabled\_for\_marketing\_drip and opted\_in\_to\_mailing\_list are positive predictors of adoption.

**Figure 1.** *Spearman correlations for target (adoption)-to-feature relationships and feature-to-feature relationships.*



**Table 1.** *Features used in models.*

<class 'pandas.core.frame.DataFrame'>

Int64Index: 12000 entries, 1 to 12000

Data columns (total 9 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 opted\_in\_to\_mailing\_list 12000 non-null int64

1 enabled\_for\_marketing\_drip 12000 non-null int64

2 GUEST\_INVITE 12000 non-null uint8

3 ORG\_INVITE 12000 non-null uint8

4 PERSONAL\_PROJECTS 12000 non-null uint8

5 SIGNUP 12000 non-null uint8

6 org\_id\_large\_org 12000 non-null int32

7 was\_invited\_by\_a\_user 12000 non-null bool

8 creation\_year 12000 non-null int64

dtypes: bool(1), int32(1), int64(3), uint8(4)

memory usage: 738.5 KB

**Table 2.** *Model Metrics.*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Best parameters | Best CV ROC\_AUC | Best Test Accuracy |
| **ExtraTreesClassifier** | **n\_estimators=25, max\_depth=None, bootstrap=True, max\_samples=0.85** | **0.6322** | **0.5925** |
| RandomForestClassifier | n\_estimators=50, max\_depth=8, bootstrap=True, max\_samples=0.85 | 0.6321 | 0.5905 |
| GradientBoostingClassifier | n\_estimators=300, subsample=0.6, loss='deviance' | 0.6279 | 0.5865 |

**Figure 2.** *Feature importances for best model*

