ECO 395 Project: Acquire More "kisses" on a Dating App

Jipeng Cheng, Weidi Hou, Yu-Ting Huang

5/8/2022

Contents

1 Abstract

We utilize a small dataset containing the information of part of the Lovoo's users to figure out who is using the dating app and is more charming than others. Exploratory analysis suggests a correlation between lovability and features such as speaking French and being mysterious. Market segmentation with PCA and clustering indicates the existence of 3 groups of users: extraordinaire, normal, and frigid users. Prediction models established by supervised learning further show that users who 1) provide more details in their profile, 2) post more pictures, 3) are more mature, and 4) are newcomers are causally more attractive. Due to the limitation of the small sample with fewer features, the results here only apply to female users and are expected to be improved with a larger dataset.

2 Introduction

The importance and social influence of dating apps are rising more and more today. For example, "The Tinder Swindler" is one of the most famous movies in 2022; it received 45.8 million hours of global views in its first week of release and hit the top 10 in 92 countries on Netflix. Furthermore, the application "Tinder - Dating New People" is super popular in the US and enables over 55 billion matches. Of course, except for Tinder, there are so many dating apps such as OkCupid, Bumble, and Coffee Meets Bagel to name but a few; they employ different features to attract different groups.

We analyzed a dataset from Lovoo, an online dating app in this study. The purpose of this research is to profile users by clustering, build a predictive model and point out what features and specific factors can acquire more "likes" (called "kisses" on the Lovoo), which serve as an indicator of people's potential charms. On the other hand, as we know that more "likes" means the person may get more matches and more

potential encounters. It is worth noting that our data only collect samples from female users. Therefore, readers may only apply the results to female users on this kind of platform.

Note that we would not actually use "kisses" as our label. Instead, we created a new label called "conversion rate", which helps avoid the problem that the number of likes correlates highly with the number of profile visits by other users, which can be polluted by information like how much time they spent on this app, compared with using the original `counts_kisses` as the dependent variable.

$$\texttt{conversion} = \frac{counts\_kisses}{counts\_profileVisits} \quad (1)$$

In the following analysis, the consequence of this project can help many young people who would like to make friends or have a pair, including our friends, classmates, and families, to employ more wise strategies when using Lovoo and other dating apps.

3 Data and Methods

3.1 Codebook

The original data comes from "Dating App User Profiles' stats - Lovoo v3" gathered during spring April and May 2015. The IOS version of the Lovoo app was in version 3 at that time. The original data includes 2940 rows and 39 variables, and we use only 22 of these features (and create 1 new label as mentioned before).

The details of each variable present in the following table.

Table 1: Variable Descriptions

Variable	Description
age	user age
counts_kisses	Number of unique user accounts that "liked" (called "kiss" on the platform) this user account
conversion	Index for converting numbers of profile visits to likes, constructed as in formula (1)
counts_details	The degree of account completion
counts_pictures	Number of pictures on the user's profile
counts_profileVisits	Number of clicks on this user (to see his/her full profile) from other user accounts
flirtInterests_chat	1 if the user indicated being in search for people to chat with
flirtInterests_friend	1 if the user indicated being open to making friends
flirtInterests_date	1 if the user indicated being open to dating people
isVip	1 if the user is VIP (this status came with benefits)
isVerified	Whether the user's account was verified through one of the methods (Facebook, phone number, ...)
lang_count	Number of languages the user knows
lang_fr	1 if the user can speak French
lang_en	1 if the user can speak English
lang_de	1 if the user can speak German
lang_it	1 if the user can speak Italian
lang_es	1 if the user can speak Spanish
lang_pt	1 if the user can speak Portuguese
freshman	1 if the user register no more than one month
hasBirthday	1 if the user has birthday
highlighted	1 if the user's profile is currently highlighted (at fetch time)

3.2 Unsupervised Leanring for Market Segmentation

We would employ some unsupervised learning techniques to divide users into different groups, which could give us informative user profiles. Given that the data contains both continuous and binary features, we would apply a generalized version principal component analysis (PCA) method called **PCAmix** for dimension reduction. More specifically, PCAmix imposes standard PCA on quantitative features and multiple correspondence analysis (MCA) on qualitative features. This allows us to perform clustering techniques like **Kmeans++** on the principal components that PCAmix provided us.
Supervised Learning Model for Prediction and Causal Inference

In order to analyze what features will help app users attract more romance, we decide to employ **LASSO regression** and **Random Forest** methods. As we all know, LASSO regression as a linear regression will give us more robust and understandable coefficients. Random forest method, on the other hand, will give us better fits but less interpretability. A complementary approach would be combining these two methods and comparing their results, which is what we are going to proceed with.

4 Results

4.1 Exploratory Data Analysis and Unsupervised Learning

4.1.1 Visualization

We would like to start with Figure 1 to extract some key information from the dataset. The top panel shows that multilingual do not prevail over other monolinguals; on the other hand, speaking French only seems very "hot". Many French speakers are very efficient in converting profile visits to likes with few pictures, and do not show clear intention about whether they want a chat, friend, or date. Mystery makes them extraordinaire! Note that in general, users with unclear intentions (like `None`) would post fewer photos on their profiles, who might want to keep mysterious (as those French speakers do!), protect their privacy, or even happen to be not passionate enough, while users with mediate intentions (like `Chat` and `Friend`) tend to post more pictures than those with strong intentions (`Date`). This suggests that making friends might require more sincerity.

The bottom panel of Figure 1 provides more demographic characteristics of the observations. Iberian (Spanish and Portuguese) speakers are more straightforward about their desire for a date, while Italian is shyer about the same thing. There is a larger portion of French speakers (as well as those multilingual) are young users, and the converse happens to speakers of other languages. These fancy results, nevertheless, echo some stereotypes.

4.1.2 PCAmix and Clustering with Kmeans++

After taking a first impression on Lovoo's users from data visualization, we would like to dive into the segmentation of these users. A powerful and conventional approach would be running clustering algorithms after conducting PCA, which is expected to improve the performance of clustering because more noises are ignored. Thus, we first run **PCAmix** on the dataset (without labels like `counts_kisses`, `counts_profileVisits` and `conversion`) and consider the first 4 principal components (`PC1`,`PC2`,`PC3`,`PC4`) given the drop of the 4th principal component's contribution to explaining the total variations. Especially, `PC1` contrasts the scores of `lang_en` and `lang_es` together with `lang_count`, `PC2` contrasts `lang_fr and lang_de`, `PC3` indicates the accessibility of a user's information (relevant to `freshman`, `counts_pictures`, `isVerified`), and `PC4` captures user's intentions. Then we try to pick an "optimal" number of clusters, i.e. k; though calculation suggests that $k = 4$ maximizes the CH index, $k = 3$ does make more sense. Performing **Kmeans++** with $k = 3$ on the 4 principal components gives Figure 2.

The left panel of Figure 2 illustrates sharp contrasts between cluster 1 and cluster 3. Cluster 3 should refer to users who have extraordinary charms (maybe those French speakers) in the sense that they are surprisingly attractive to profile visitors so that visitors are really likely to like them. On the contrary, cluster 1 represents normal people since they need a lot more visits to have the same number of "kisses" as cluster 3; in a word, their search cost of dating can be higher! The right panel indicates the existence of

Figure 1: Value of features for wines of different colors

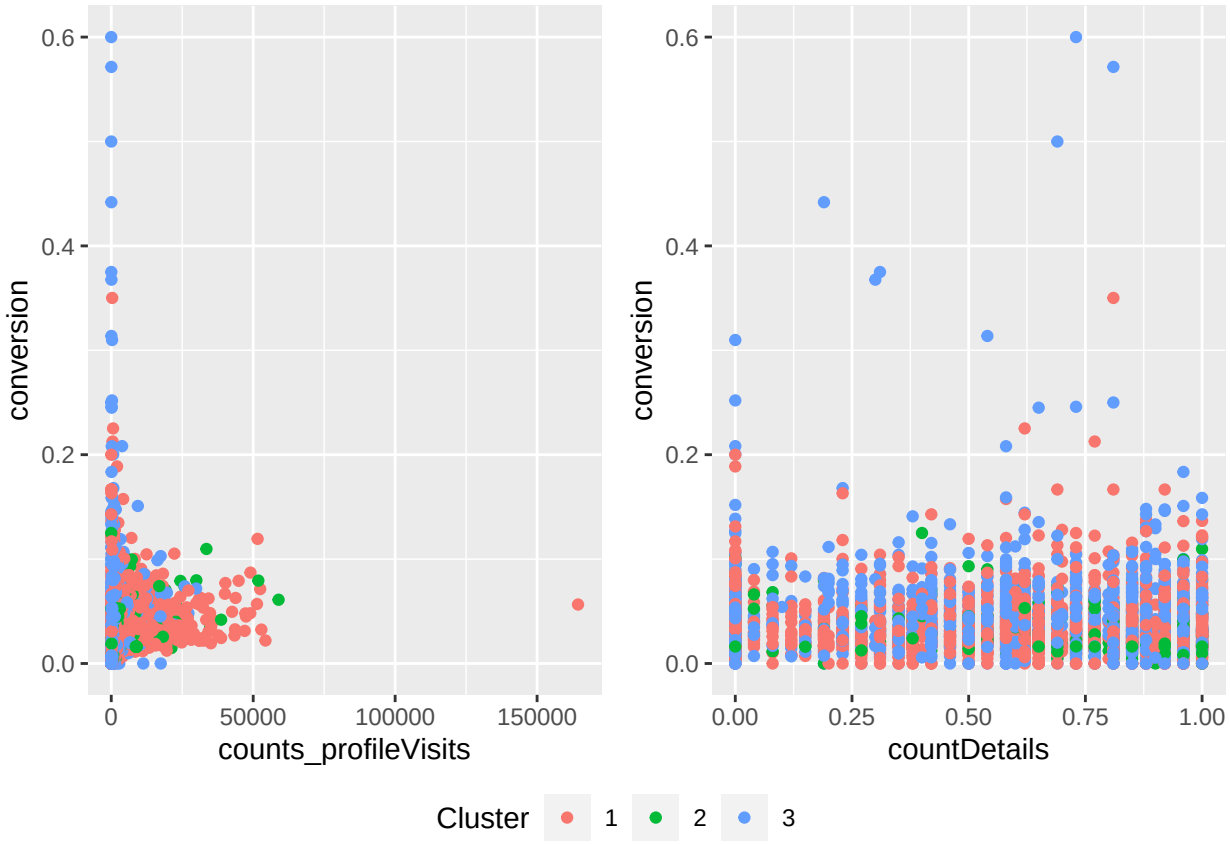

Figure 2: Value of features for wines of different colors

cluster 2, which has a frigid attitude to dating because they really do not care about their profiles (and do not post much information). Their conversion rate, as expected, is lower on average than the other 2 clusters.

The outcome of unsupervised learning describes the in-sample user portraits. Next, we will tackle the real problem: what determines your attractiveness on a dating app?

4.2 Supervised Learning

4.2.1 LASSO

LASSO approach is to make some regularization so that the regularized fit minimizes the deviance plus a **penalty** on the complexity of the estimate:

$$\min_{\beta \in R} dev(\beta)/n + \lambda \times pen(\lambda) \text{ (2)}.$$

Here λ is the penalty weight, while **pen** is some cost function that penalizes departures of the fitted β from 0. In order to use **gamlr** function in r programming, we need to create our own numeric feature matrix.

4.2.2 Cross-validated LASSO

Then we use cross-validated LASSO regression method so that the result will be more robust.

Table 2: LASSO Coefficients

Variable	beta_hat
intercept	-3.457606879
age	0.013822836
counts_pictures	0.004166058
lang_fr	0.040604945
lang_de	-0.127865953

Notice that beta coefficient for the following variables including `flirtInterests_chat`, `flirtInterests_friends`, `flirtInterests_date`, `isVIP`, `isVerified`, `lang_en`, `lang_it`, `lang_es`, `lang_pt`, `countDetails`, `freshman`, `hasBirthday`, and `highlighted` are **zero**.

The result shows that the optimal $log\lambda$ which minimize the test set mean square error is -8.99 and under this level of penalty, the fitted coefficients that are not zero are `age`, `counts_pictures`, `lang_fr`, `lang_de`. From these results, we conclude that age, number of pictures on the users' profile and proficiency in French have a positive effect on being loved by more people. However, proficiency in German has a negative effect on their level of charm. These results tell us that being mature will help people attract more people and more pictures on their app profile will also help them become popular since more pictures will tell other people more information about themselves and help other people know the user better. In addition, people who have a good looking or who are more out-going and more confident prefer to post their picture on their app profile. These type of users also attract more people. Besides, French people are more romantic and they have a higher probability to attract more people on the app. On the other hand, German people are more serious and introverted, that's why the fitted coefficient of knowing German is negative. In addition, from our regression result, we found that our feature matrix is very sparse (i.e, mostly zero), this is especially true since we have lots of factors as features. LASSO regression is a great way to improve the efficiency of our model since it screens and ignores zero elements in actually storing X.

4.2.3 Random Forest

A random forest starts from bagging. We still take B bootstrapped samples of the original data and fit a tree to each one, and we still average the predictions of the B different trees. However, it adds more randomness. With each bootstrapped sample, we don't reach over all the features in x when we do our greedy build of a big tree. Instead, we randomly choose a subset of a features sub sample to use in building that tree. The

advantages of using fewer features in each tree are that it simplifies each tree, reducing its variance and it diversifies the B trees, and decorrelating their predictions.

From the load.forest plot, we found that more trees shows smaller out-off sample MSE. After 500 tress, the partial decresing of MSE becomes very small.

Then we study how random forests can give us a variable importance measure by Figure 3.

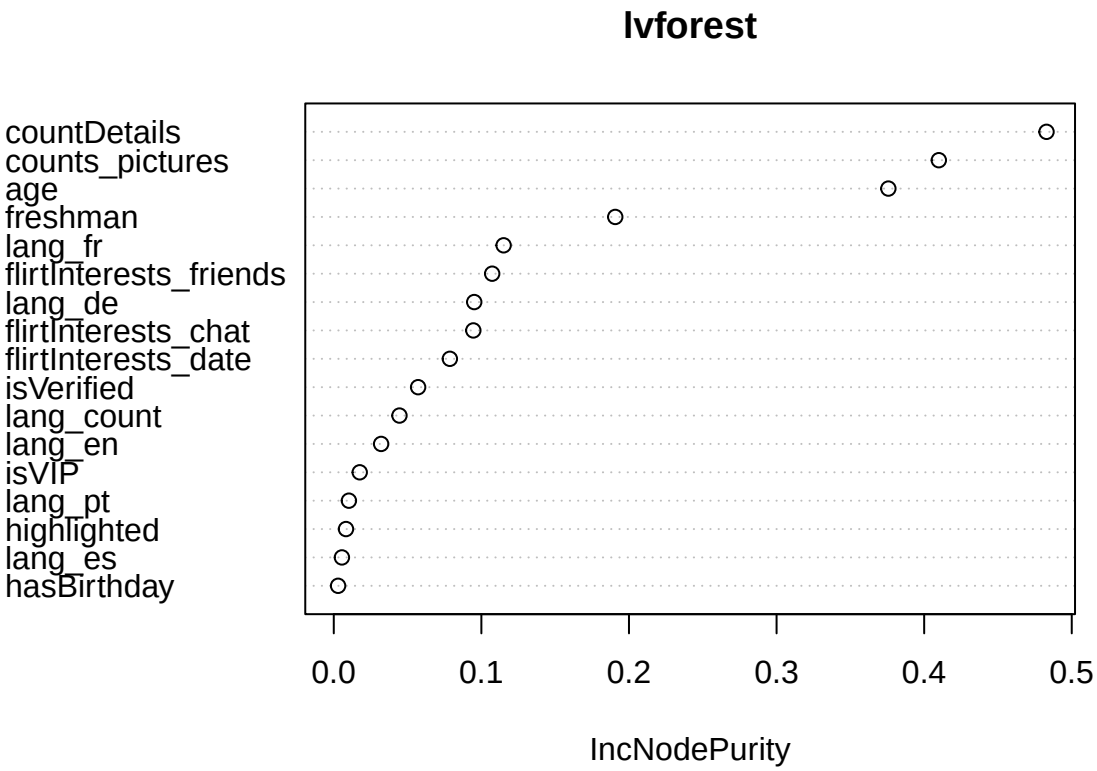

Figure 3: Variance Importance Measures

The x axis represents percentage increase in mean square error and the y axis represents different variables. The variable importance plot shows how much omitting each of these variables inflates the MSE of the prediction, higher is worse. From our result plot, we found that `counts_Details`, `counts_pictures`, `age`, `freshman`, `lang_fr` and `lang_de` have higher value, which means these variables are more important in this model, they have higher influence on whether an app user will attract more people. This result is rational since completed information on their profile and more pictures on users' profile will provide more information of the user. In addition, users' basic information such as their age and whether they are new users will also provide some information their background and their communication proficiency, which are also key factors deciding whether they can attract more people. Finally, people's language will tell us where they come from. People from different country may have different characteristics and their social development environment will also impact their probability of attracting more people.

Then we study the partial importance of each variable by eyeballing Figure 4.

First, from the PD on `age`, we found that when people above 20 years old, they are more likely to attract people, especially when people are above 25 years old, the slope of `age` is much higher above 25. It's reasonable since age 20-26 is a period of dating and getting marriage so people during this age are more likely to log onto this app to find their dating mate. Second, the result of PD on `counts_pictures` and on `counts_Details`, the result are rational since more pictures and more completed profile will tell people more stories about the users themselves, which have a positive effect on attracting more people. Third, from language perspective, people speaking French are more likely to attract people but people speaking German

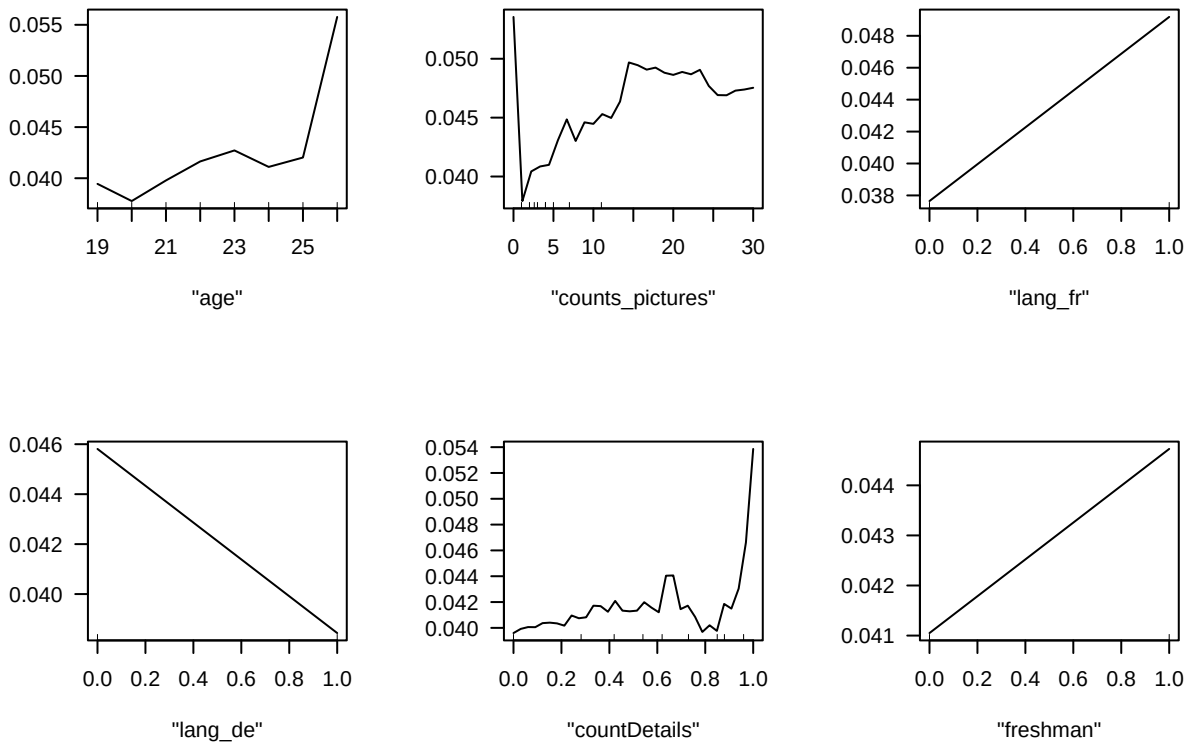

Figure 4: Partial Dependence of Key Variables

are less likely to attract people, since people's characteristic are influenced by their nationality and their growing up environment. Finally, from the PD of the `freshman`, it seems that new user are more attractive, since people always like new things.

5 Conclusion

We utilize the information of part of the Lovoo's users to figure out who is using the dating app and what users can be more attractive. Exploratory analysis suggests a correlation between lovability and features such as speaking French and being mysterious. Market segmentation with PCA and clustering indicates the existence of 3 groups of users: extraordinaire, normal, and frigid users. Prediction models established by supervised learning further show that users who 1) provide more details in their profile, 2) post more pictures, 3) are more mature, and 4) are newcomers are causally more attractive. Due to the limitation of the small sample with fewer features, the results here only apply to female users and are expected to be improved with a larger dataset.

6 Reference

1. Tinder, https://tinder.com
2. 'The Tinder Swindler' Becomes First Doc To Lead Netflix's Weekly Film Chart, https://deadline.com/2022/02/the-tinder-swindler-first-doc-lead-netflixs-weekly-film-chart-1234928573/
3. Lovoo, https://about.lovoo.com/en/#app-features