Shortcut to love: an analysis of dating app user profiles

Introduction

Dating applications and websites have experienced an unprecedented surge in user numbers during Covid-19. According to Tinder's 2020 report, more than 3 billion users 'swiped' on March 29th, 2020 – a record 130 times higher than the previous year (*The Future of Dating Is Fluid*, 2022). Other dating apps also saw similar growth during the lockdown period in 2020. Between March to May 2020, dating applications like OkCupid and Bumble have seen 700% more dates and a 70% increase in video calls respectively (Robison, 2020). Dating applications provide users with virtual interactions and connections that were absent during lockdown, but it is still very different to dating offline. The nature of virtual dating makes a person's profile the only source of information that others can make judgments from. Apart from photos, a profile consists of information like gender, age, location, and a few lines of self-introduction. This form of dating provokes an interesting research question: what kind of features in a profile will attract the most people?

This study will explore this question using a dataset[1] collected from various users on 'Lovoo' – a dating application based in Germany with 30 million active users worldwide. The original dataset has 3992 rows and 42 columns, with 366 missing values. However, because the algorithm of the application recommends only one's preferred gender, in this case, only female users' data has been collected. Hence this study is based on 3626 female application users who live in Europe with ages ranging from 18 to 28.

This study aims to find the relationship between the number of visits to a profile and other variables like age, number of pictures, number of languages spoken, and whether the user is open to making friends, having chats, or dating other users. To do this, this study makes use of models trained by two supervised machine learning tools – a classification tree and a feedforward neural network. The models' predicting power is compared using the test set and compared in terms of performance.

[1] https://www.kaggle.com/jmmvutu/dating-app-lovoo-user-profiles

There are several assumptions in this study. Since the data was collected from the male perspective, it is limited to viewing and gaining data from female profiles. The profile's views therefore come from that user's 'genderLooking' setting – either the views can come from males only, or both males and females. The data therefore does not include females looking solely for other females, and there is no data on male profiles. Moreover, this study excludes the impact of physical appearance, and it also assumes more visits equate to greater chances of getting matches, whilst the number of visits also reflects one's popularity on the app.

The mean number of visits is 3705, hence we create a response variable named 'High' which assigns profiles that have been visited more than 3705 times as 'Yes', otherwise 'No'. This study uses 7 exploratory variables: genderLooking, age, counts_pictures, lang_count, flirtInterests_chat, flirtInterests_friends, flirtInterests_date, and one response variable High, which is a feature created based on the summary of the number of visits. It also removes rows that have 'none' in genderLooking for higher accuracy.

Table 1 - Summary of visits

Min. 1st	Qu.	Median	Mean	3rd Qu.	Max.
0	383	1222	3705	4063	164425

Table 2 - The 7 variables in the dataset

genderLooking: 1 for male, 0 for both

age: 18-28

counts_pictures (number of pictures visible in the profile): 0-30

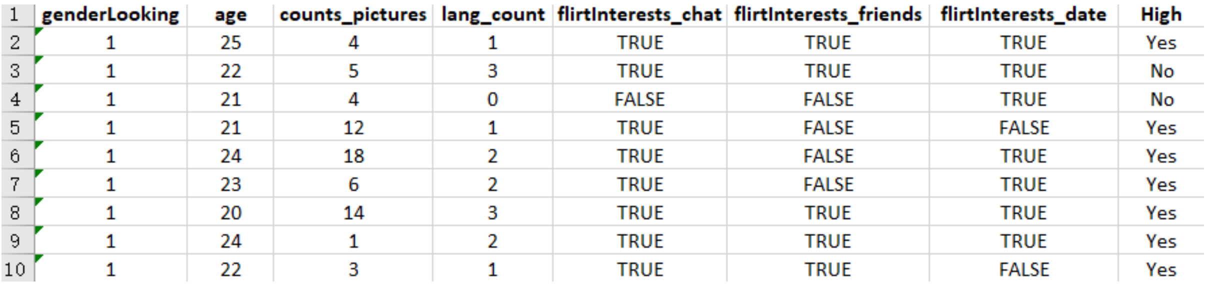
lang_count (number of languages a person can speak): 0-9

flirtInterests_chat (openness to chatting to other app users): TRUE or False

flirtInterests_friends (openness to making friends with other app users): TRUE or False

flirtInterests_date (openness to dating other app users): TRUE or False

High: Yes (visits >= 3705), No (visits<3705)



1	genderLooking	age	counts_pictures	lang_count	flirtInterests_chat	flirtInterests_friends	flirtInterests_date	High
2	1	25	4	1	TRUE	TRUE	TRUE	Yes
3	1	22	5	3	TRUE	TRUE	TRUE	No
4	1	21	4	0	FALSE	FALSE	TRUE	No
5	1	21	12	1	TRUE	FALSE	FALSE	Yes
6	1	24	18	2	TRUE	FALSE	TRUE	Yes
7	1	23	6	2	TRUE	FALSE	TRUE	Yes
8	1	20	14	3	TRUE	TRUE	TRUE	Yes
9	1	24	1	2	TRUE	TRUE	TRUE	Yes
10	1	22	3	1	TRUE	TRUE	FALSE	Yes

Figure 1 - Sample image of dataset used in this study after cleaning

Modeling process

Classification Tree

This model builds a classification tree by first randomly splitting the data into 80% training and 20% testing datasets, corresponding to 2900 and 726 observations respectively. The input is seven variables, namely, genderLooking, age, counts_pictures, lang_count, flirtInterests_chat, flirtInterests_friends, flirtInterests_date, and the model output is a binary class High (yes or no). The classification tree model is implemented through the Rpart (recursive partitioning) package in R, which is generally considered to be more flexible than the tree() function, as this method involves more parameters during modelling (Chauhan, 2018). The split in Rpart is chosen by recursively trimming off least important splits in accordance to the complexity parameter (cp) and it uses Gini Index as the class purity metric (Comprehensive R Archive Network (CRAN), 2022).

After fitting a classification tree to training data using Rpart (), we use the 10-fold cross-validation to cross-validate the tree. This classification tree model has 4 splits with one root node and 4 terminal nodes. As this tree does not have an excessive number of branches, it is relatively simple, thus no pruning is operated here.

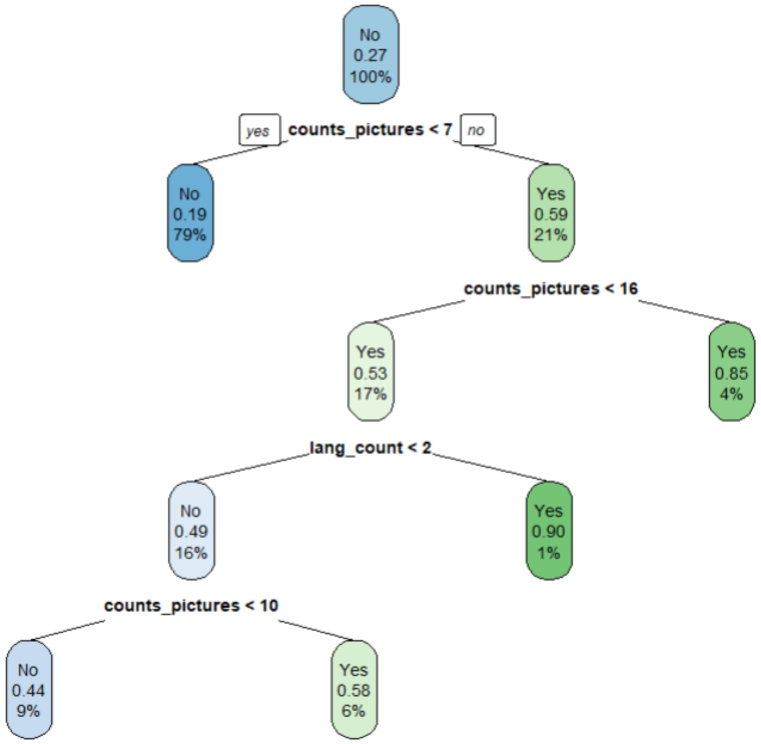


Figure 2 - Classification tree model

In this tree model, each node contains three pieces of information: the predicted class (Yes or No), the likelihood of a profile with this classification having a high number of visits, and the percentage of observations in the node. The first node shows only 27% of users in the entire dataset have over 3705 visits in their profile. We also find that 79% of users have less than or equal to 7 pictures in their profiles, and in this case, they will have only a 19% chance of having a popular profile. However, if one owns more than 7 pictures, this user is 59% more likely to have a popular profile. The next split points out again the positive correlation between the number of pictures and visits. If a female user has more than 17 pictures on her profile, she has an 85% percent chance of having a profile with a high number of visits. Meanwhile, for those with fewer than 16 pictures, if they speak more than 2 languages, they are also 90% more likely to have a high number of profile visits using this dating app. Moreover, for monolingual people, uploading more than 10 pictures could give them a 58% chance of becoming more popular.

These findings make sense: people may have a stronger connection with users who share more pictures and therefore more information about their appearance and lifestyle. It also suggests female users who share less about themselves through pictures but are bilingual are more attractive to other users than monolinguists, which could be because they are more likely to be considered educated or open-minded, or because they have an increased chance of being able to communicate with users of different languages.

In terms of feature importance, variable counts_pictures appear 3 times and lang_count only once in the classification tree. According to the feature importance table, counts_pictures, lang_count, and genderLooking are the most important features in this study, however, number of pictures greatly outweigh the other two variables. Since genderLooking has a score of only 0.3 it is not used in building the tree. The absence of ‘flirtInterests_chat’, ‘flirtInterests_friends’, ‘flirtInterests_date’ indicate these variables are not chosen during the splitting process, and therefore have the least contribution to visit counts.

Table 3 - Variable importance table

counts_pictures	lang_count	genderLooking
169.7049179	13.0082055	0.3097192

Next, the classification tree is fitted to the 726 observations in the test set to make predictions as to which profiles will have a high number of visits, which is compared with the actual response for profile visits in the testing data. This predicts 645 users to be in the less popular category and 81 in the popular category, among which 581 are in the correct group giving a prediction rate of 80%, i.e. this proves the model can have an 80% accuracy in predicting if a user’s profile can get a high number of visits.

Table 4 - Confusion matrix of model 1

Prediction	No	Yes
No	519	126
Yes	19	62

Feedforward Neural Network

The second model employs a neural network and creates a feedforward neural network model using Keras. The advantage of using a neural network means it can increase model accuracy with time, at the same time analysing data at a fast speed (Education, 2021). As a basic form of deep learning model, the feedforward neural network is flexible to use, as it does not require us to find the precise parameters, instead only specifying the right general function (Goodfellow et al., 2016).

The neural network model has three components: input layer, hidden layer, and output layer. Each neuron in the hidden layer takes in a linear model with weighted inputs and bias, and then the sum will go through an activation function (Techopedia, 2018), finally, it will provide a binary output. This process is also one-directional. Since we aim to solve a binary classification problem as well as performing prediction, the input here is seven feature variables as in the previous section. Besides, because the output in this study is a binary value indicating either high or low visits counts, we use 'sigmoid' as the activation function in the last layer of hidden layer.

In order to access more stable data in the feedforward model, both input and output variables are normalised and scaled to a range [0, 1]. Binary variables with answers like true or false, yes or no are converted into dummy variables (1: True/Yes, 0: False/No). After scaling, the input data becomes a 2-dimensional tensor with 3626 rows and 7 columns, and our output is a 1-dimensional array between 0 or 1. To train the model, we first feed it with 80% of the original data and test its prediction rate on 20% testing data, randomly selected.

After defining feedforward as the model architecture, we use Keras.Sequential() function. In constructing the model, three layers are added to avoid overfitting the data. The first layer reshapes and flattens the 3626 x 7 array into a 25382-length vector. The second layer takes in 14 neurons and use 'relu' as the activation function and the last layer takes in 1 neuron using 'sigmoid' as the activation function. Next, the model is compiled with RMSprop as the optimiser. For the loss function, we choose binary cross entropy, and we use accuracy as the metric. The model is then fitted to the training data with batch sizes of 30 and 50 epochs and is then evaluated on the testing set, as with smaller batch sizes and more epochs one can increase the final test accuracy despite increasing the runtime.

Table 5 - Summary of the feedforward neural network model

Layer(type)	Output Shape	Parameters
flatten	(None, 7)	0
dense	(None, 14)	112
dense_1	(None, 1)	15

The final model has 127 parameters with a test accuracy of 76% with 51% loss, which means this model has a 76% likelihood of correctly predicting a profile's popularity. However, although the loss is lowered each epoch from 69%, the final test loss is still high, which suggests that the predicted and expected values are very different, and the test-error is high. This means that this model is not as predictive for the problem as model 1.

Table 6 - Loss and accuracy table of training set and test set

Training loss: 0.4758	Training accuracy: 0.7861
Test loss: 0.5112	Test accuracy: 0.7546

Table 7 - Confusion matrix of model 2

Prediction	No	Yes
No	482	29
Yes	158	57

Discussion

From the prediction performance gathered above, it is evident that classification tree is the better model. First, it is straightforward and interpretable, allowing us to have a better understanding of the relationship between profile visit number and other features, while seeing the importance of each variable. Feedforward neural network, on the other hand, does not provide feature importance, hence we have only the prediction accuracy. Although both models have a relatively high prediction accuracy, neural network suggests a weaker performance as the loss of the network is very high. The feedforward model itself is far less interpretable compared with the decision tree as one can’t explain what each hidden layer does to the overall model.

Based on the predictions of the decision tree model, an average female user on Lovoo has a 27% probability of gaining more than 3705 visits on their profile. Having more pictures definitely brings more visits to profiles, compared to any effects of the users age, or preferences in terms of looking to date, chat of make friends. However, if a user has less than 16 pictures, it seems that being able to speak more than 2 languages can also be an attractive characteristic for increasing visits.

But why are pictures important to a dating app profile? Apart from representing physical attractiveness, it has been argued that pictures also suggest that a person is ready to go offline and meet someone, and they are serious about finding their other half (Healthy Framework, 2020). However, lighting, poses, facial expressions, environment, etc. in the photos can also ‘make or break’ a potential match (Johnson, 2022) – which is not something considered in our models. Being bilingual, on the other hand, is considered as a charismatic advantage to most people. According to a survey, an ‘overwhelmingly majority’ of people think the ability to speak more than one language is an attractive feature (Owen, 2016).

This study also shows that a simple model like a decision tree can sometimes work better than the advanced deep learning method. This is not coincidental, as others have pointed out that tree-based methods ‘routinely outperform’ neural network models. This is because in building a decision tree, only features that create the best split are kept and the rest are not used (*Decision Trees vs. Neural Networks*, 2012). Therefore, by eliminating less important variables, decision trees can achieve interpretable models that also have high predictive accuracy. This argument is supported by this study as well, as only 2 out of 7 features are used in the final tree model. On the other hand, a basic feedforward neural network uses every feature in building the model without selection, which might not be the best solution for every problem. It is claimed that data which does not require ‘intermediate representation’ (*[D] Why Neural Networks for Tabular Data Are Bad?*, 2021) is more suitable for trees as opposed to neural network, whereas neural networks fit better with large uninterpretable data, like images or sound.

There are however some limitations to this study. One key parameter that is not available from the data is the length of time the user’s profile has been active. It is almost certain that a profile that has just been created, despite perhaps having a large number of photos, will have less visits than a profile that has existed for a longer time. This therefore would have been a significant source of error for both models. The other limitation to both models is in determining popularity via a binary assignment: assigning a profile as either a high or low number of visits doesn’t fairly represent a large number of profiles near to the mean, i.e. a profile with 3706 visits is considered to have a high popularity, whereas a profile with 3705 visits will have a low popularity. One could increase the fidelity of the models by having more classes for popularity.

The conclusions of this study open up more research questions to be answered in future works: If the data is not limited to only female users, we can compare the preferences of male and female users, to find out if there are common features that both groups prioritise. Moreover, if data with pictures was provided, we could then analyse physical features using convolutional neural networks. We could also explore whether users prefer ‘selfies’ or photos from a third-person view, and whether photos that have specific environments e.g outdoors/indoors, or additional features such as pets will achieve more visits. Answers to these questions may further our understanding of the behaviours of users when seeking partners online.

Executive Summary: an analysis of dating app user profiles

Background

Many dating applications saw unprecedented user growth during the Covid-19 pandemic, especially 'Tinder', 'OkCupid' and 'Bumble'. Many use them to interact with others and find partners in the virtual world. However, a popular profile is vital to being able to find successful matches.

Aim

This report aims to analyse which features of a profile has the biggest impact on the number of visits. It has three objectives: first, building models and understanding which variables play the biggest part in determining the number of profile visits. Second, making predictions of a profiles popularity based on the models. Third, to compare the performance and limitations of the models.

Data

This report uses a dataset collected from the information of 3626 female users on a German dating application named 'Lovoo'. These users all live in Europe and are aged between 18-28.

Method

We consider 7 variables: the number of profile pictures, age, preferred gender, languages spoken, and the preference of making friends, chatting or dating with other online users. We create two models, one based on a decision tree, and the other using a neural network. We train both models using 80% of the data, and test the model's predicting power against the remaining 20%.

Conclusions

The decision tree model shows that if a user has more than 10 photos, they will have a 58% chance of getting more than 3705 visits in their profile. But if a user has fewer than 16 pictures, if she can speak two or more languages, she can have a 90% chance of having a high number of visits. This model gave a prediction rate of 80%. The second method uses a neural network and it chooses the same 7 variables to train a deep learning model. Despite its computational complexity, this model only has a prediction rate of 76%, meaning it is less accurate than the previous model. It also gives no indication of the importance of each variable.

This study finds that the classification tree model applys better to tabular data and is more interpretable, and it is more suitable for problems in which the raw data is relatively simple to understand. In contrast, neural network models apply better to complex data or those that require intermediate interpretation in order to be understood.

For someone looking to find connections via a dating app, the study shows that increasing the number of photos available to be viewed, and the number of languages spoken will increase a user's popularity. More work can be done however in considering additional variables not included in this dataset.

References

Chauhan, R. S. (2018, June 13). *A comparison on using R.Tree vs R.Rpart – rohitschauhan*. Http://Www.Rohitschauhan.Com/. http://www.rohitschauhan.com/index.php/2018/06/13/a-comparison-on-using-r-tree-vs-r-rpart/

Comprehensive R Archive Network (CRAN). (2022, January 24). *CRAN - Package rpart*. Https://Cran.r-Project.Org. https://cran.r-project.org/web/packages/rpart/

[D] Why Neural Networks for tabular data are bad? (2021, March 7). Reddit. https://www.reddit.com/r/MachineLearning/comments/lzoqjg/d_why_neural_networks_for_tabular_data_are_bad/

Decision trees vs. Neural Networks. (2012, July 17). Software Engineering Stack Exchange. https://softwareengineering.stackexchange.com/questions/157324/decision-trees-vs-neural-networks

Education, I. C. (2021, August 3). *Neural Networks*. Https://Www.Ibm.Com. https://www.ibm.com/cloud/learn/neural-networks

The Future of Dating Is Fluid. (2021). Tinder Newsroom. https://www.tinderpressroom.com/futureofdating

Goodfellow Et Al., I. (2016). *Deep Learning*. https://www.deeplearningbook.org

Healthy Framework. (2020, November 24). *Why Dating Pictures are Important for More Matches*. https://healthyframework.com/dating/blog/why-dating-pictures-are-important-for-more-matches/

Johnson, E. B. (2022, January 4). *Dating and the importance of profile pictures | Practical Growth*. Medium. https://medium.com/practical-growth/why-your-profile-picture-is-preventing-you-from-scoring-in-the-date-department-a357fe755e47

Owen, E. (2016, August 27). *Knowing Another Language Makes You More Attractive*. Travel + Leisure. https://www.travelandleisure.com/travel-tips/bilingual-people-more-attractive

Robison, C. C. A. M. (2020, December 14). *This cuffing season, it’s time to consider the privacy of dating apps*. Brookings. https://www.brookings.edu/blog/techtank/2020/11/20/this-cuffing-season-its-time-to-consider-the-privacy-of-dating-apps/

Seppala, E. (2012, February 14). *Discovering the Secrets of Long-Term Love*. Scientific American. https://www.scientificamerican.com/article/discovering-secrets-long-term-love/

Techopedia. (2018, September 5). *Hidden Layer*. Techopedia.Com. https://www.techopedia.com/definition/33264/hidden-layer-neural-networks