Project date-a-scientist

Analysis of OKCupid data

Introduction

In recent years, there has been a massive increase in the use of dating apps to find love. Many of these apps use sophisticated data science techniques to recommend possible matches to users and optimize the user experience. These apps give us access to a wealth of information we never had before about how how different people experience romance.

Goal of project :

In this portfolio project, we will analyze some data from OKCupid, an application that focuses on using multiple choice and short answers to match users. We will perform these analyses using the skills we have learned so far, including machine learning skills. We will use the information (variables) provided in the data to help us verify matches between users.

Data :

The project has one data set provided by Codecademy called profiles.csv. In the data, each row represents an OkCupid user and the columns are the responses to their user profiles which include multi-choice and short answer questions.

Analysis :

In this project we will use descriptive statistics to get an idea about the distribution of the data, but also a numerical summary of the data. We will use data visualization to see which variables are correlated and how each variable affects the correspondence between users of the site. Finally we will use supervised learning techniques, in this case classification algorithms, since the objective of the project is to make predictions about the user's astrological signs.

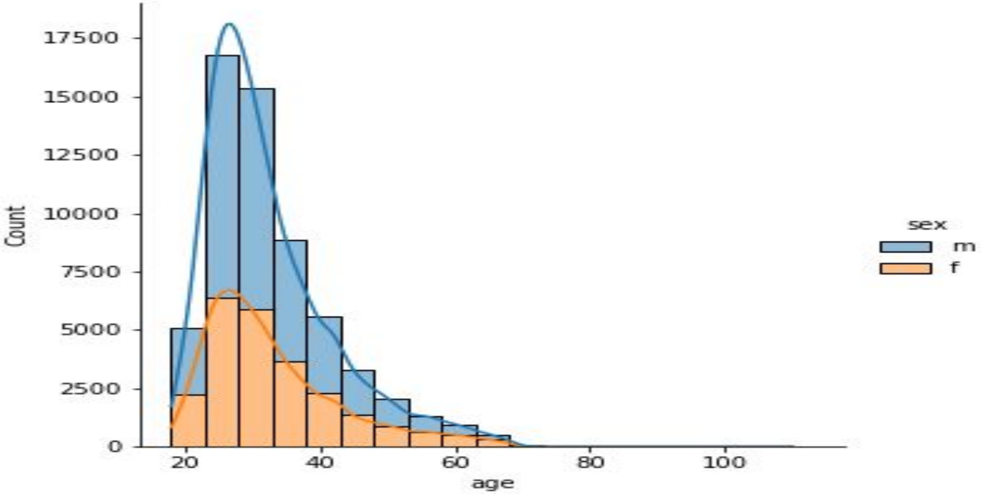
Evaluation

The project will conclude with the evaluation of the machine learning model selected with a validation data set. The output of the predictions can be checked through a confusion matrix, and metrics such as accuracy, precision, recall, F1 and Kappa scores.

Extract from the data mining section

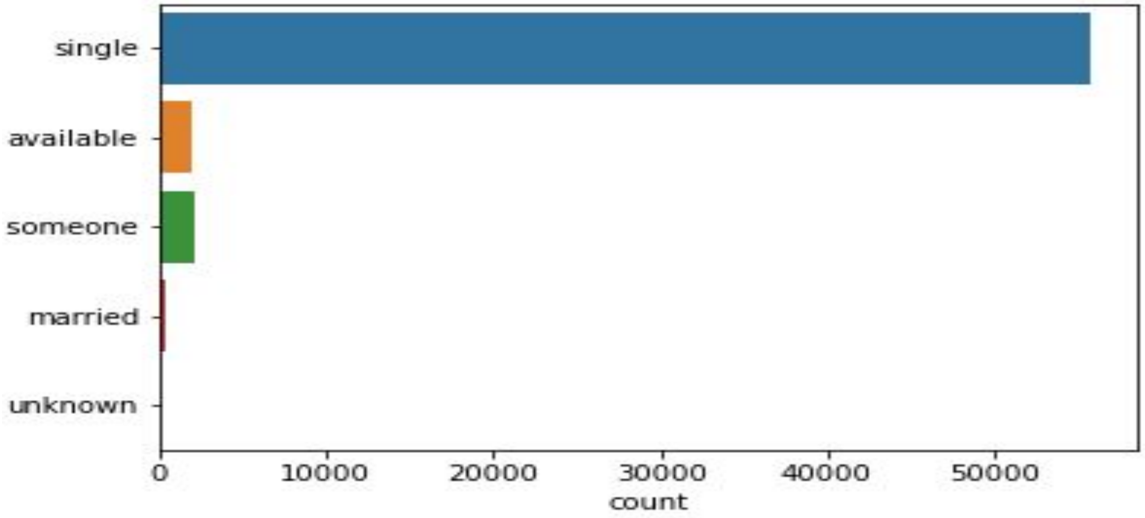
In this section we will use some of the graphics used in the project

This graph shows the variation in age by gender.



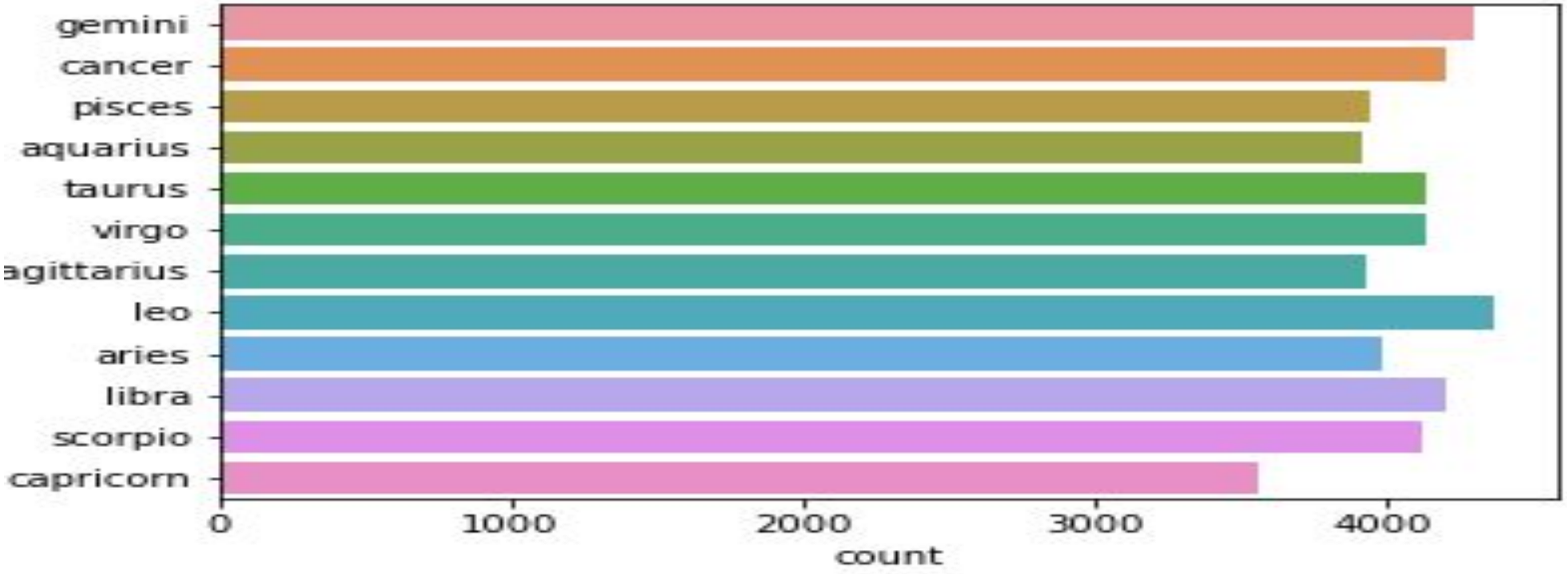
The graph shows a significant male presence relatively higher than the opposite sex.

This graph shows the status of the users of the site.



It is clear that most people who use a dating site are single. The graph below corroborates this fact.

Astrological sign



Main question

The main one that is known to be asked during the realisation of this project is the following:

What are the variables that could influence a possible match between two users?

To answer this question, we used the following variables: 'body_type', 'diet', 'orientation', 'pets', 'religionCleaned', sex', 'job', 'signsCleaned'.

Main question

To make good use of this data, it was necessary to carry out some processing on certain columns, among them: 'religionCleaned', 'signsCleaned'

The religionCleaned column has been cleaned to get clear data on religions, the same for the signsCleaned column for astrological signs.

These data allowed us to make our predictions on the machine learning algorithms we chose.

That is to say: the nearest neighbours algorithm, logistic regression, decision trees and random forests.

Some algorithms used

As our dependent variable is a categorical one, we used classification algorithms to perform the prediction.

These algorithms are :

Logistic regression

K-nearest neighbours

Decision trees and random forests

Some algorithms used

After using these different algorithms, the decision trees were the best, because they provided the best scores. More information can be found in the source code of the project.