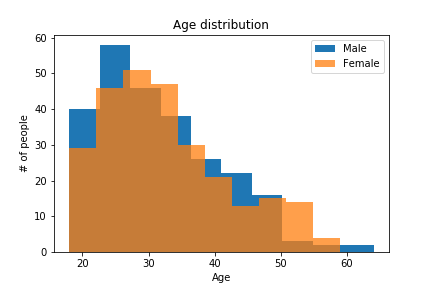
Name: Cong Do Nguyen

Assignment: Project report

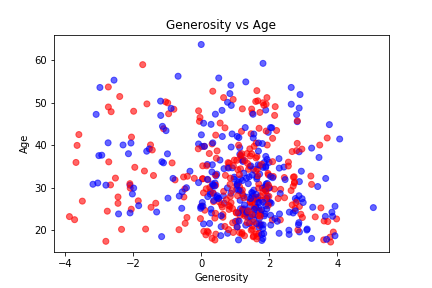
The data set given consists data of 3500 randomly chosen users from the largest Russia-language massive online social network. It has 4 main columns which consist of: social network user’s gender, their age, their profile picture grade randomly rated by the researcher’s avatar from the scale 1 to 6 (response grade), and the last columns contains the same type of grade users were given by others (stimulus grade).

Based on the data, there are 2977 missing values in the response grade column. There are 1573 male subjects and 1927 female subjects in the data. After handled missing values in the data, there are 523 clean subjects in the data left. Now there is 253 male subjects and 270 female subjects. The graph below shows the age distribution of subjects in the data. The y axis shows the quantity of subjects who have certain age corresponding to the x axis.



As we can see, most subjects from the data are from 15 to 35 years old apply to both male and females. Numbers of male subjects are slightly higher than female subjects from age 15 to 25. Above age 50, there are more female subjects than male subjects.

Generosity is the difference between the response grade and the stimulus grade. Due to the high volume of people who has the same age and they are a few discrete levels of generosity, most points in the graph of generosity and age will collide. Therefore, I slightly change the scatter points by adding some random small amount of displacement (from 0.0 to 1).



The graph above shows generosity versus age. As we can see, the data mostly dense within 0-2 range of generosity (x axis) and 15-40 of age (y axis). From this we can say that most people in the age of 20 to 40 have generosity.

In this project I produced 2 data models: logistic regression and random forest classifier. Logistic regression has the score of 0.66 while random forest classifier scores 0.97. So in this case, I would highly recommend random forest classifier as a model to make predictions.