CS-573 Data Mining HW2

Mijanur Palash (mpalash@purdue.edu)

October 7, 2020

Q2 Observations

0.1 Observations from the Bar plot

In fig-1, the red bars represent the male and green bars represent the female preferences on different attributes. We see the highest value for red occurs for"attractive_importance" attribute. So we can say for the male population looking attractive is the most important attribute they look in a potential partner. The second most important factor for them is "intelligence" of the partner which comes with relatively lower value.

The highest value for green occurs for "intelligence_important" attribute and "sincere_important" comes after that. So for an women intelligence is the top most criteria in a partner followed by sincerity. Women put "attractive_important" at the third place.

So we can say while males give much importance on attractiveness, females are more focused on the personality traits.

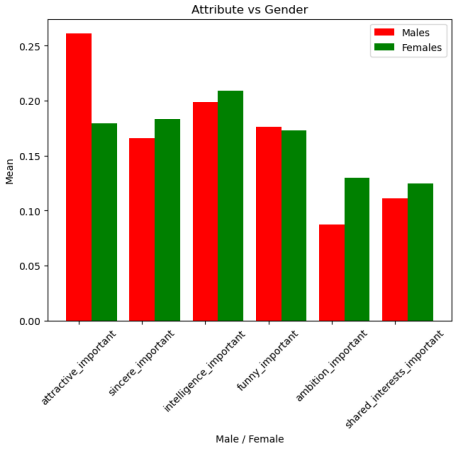
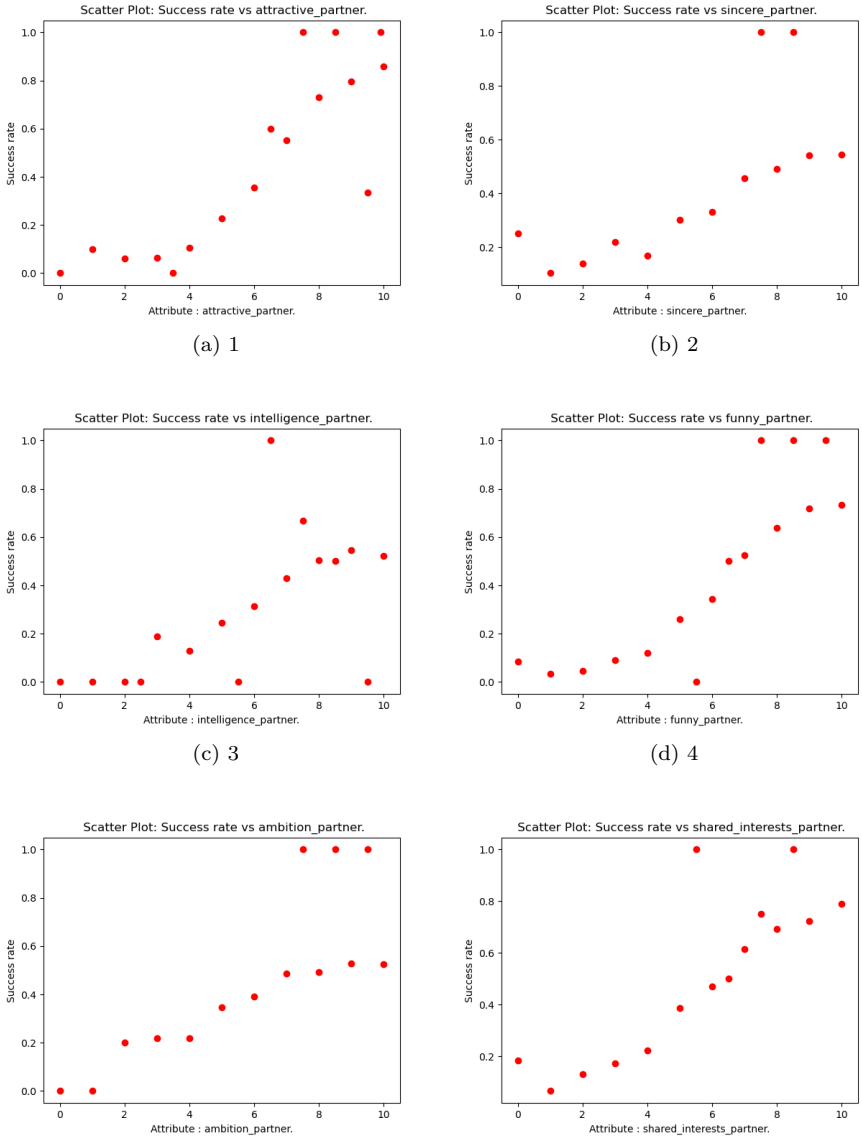


Figure 1: Males and females evaluation on the six attributes in their romantic partners.

0.2 Observations from the scatter plots



(a) 1 (b) 2

(c) 3 (d) 4

(e) 5 (f) 6

Figure 2: Values of six attributes and corresponding success rate.

From the scatter plots 1-6 of figure 2 we observe some very interesting tendency.

- people who are very attractive (7 or above rating) are highly successful to get a second date. While People who score below 5 rarely get this offer. This is probably due to the males high priority on attractiveness we saw earlier.
- Other plots also show similar properties i.e the higher rating a person get from the partner, the more likely for him to get a second date, even though the extent of the likelihood varies between the attributes.
- higher ratings in attractiveness, funny and shared interest attributes lead to higher chance of second date. So this attributes are more important than the other three.
- It is also looks like people scoring 2 or less in any attributes has very less chance for the second date.

Q5 Observations

0.3 Effects of the number of bins

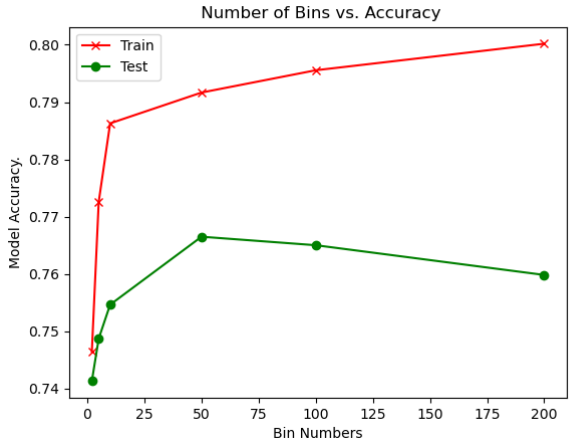


Figure 3: Number of bin vs model accuracy.

From figure 3 we observe that training accuracy keep increasing with the number of bins- very sharp at first and then slower. While the test accuracy follows the training accuracy for some time and then break out- it no more increases rather start decreasing. This thing is called over-fitting where model does not learn anymore rather memorizes the training data so training accuracy increases and test accuracy decreases. With the more bins we categorize the continuous valued data into more categories. This is initially good but when we have so many bins, there is not enough sample in each categories which leads to the overfitting.

0.4 Effects of the values of f

From figure 4 we see that with increasing values of f, the training accuracy reduces from a higher value to a lower value and at some point and beyond of f value, training accuracy looks kind of stabilized at that level.

However the test accuracy shows a different pattern. It starts with a smaller value and keep increasing with f and stabilizes within a accuracy label at the same values of f where training accuracy also stabilized.

This is because f determines the number of data points used for training purpose. When f is very small we have very few training data and the model does not learn properly. Moreover due to few training points it probably memorizes and fits very well with the training data and hence we get high training accuracy. But as this model is poorly learnt, it perform very badly in the test data which it did not saw before and couldn't memorize.

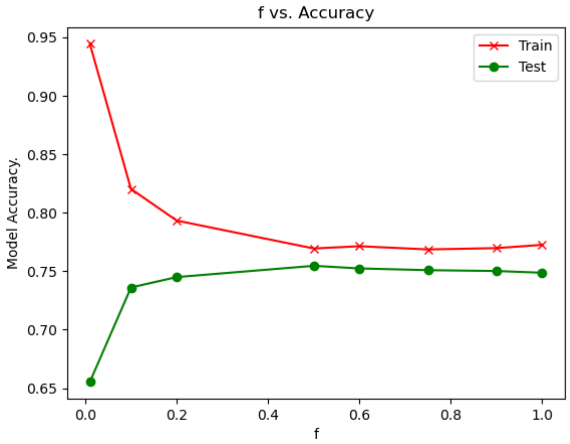


Figure 4: value of f vs model accuracy.

With f increasing, we get more and more training data points and the model learns better and better and memorizes less. So train accuracy drops and test accuracy increases-the move toward each other. At some point beyond f=0.5 the model have enough data and can't learn any more so the training and test performance kind of stabilizes. And in the higher values of f near 1 it seems like the model kind of over-fitting and the accuracy lines are kind of separating again.

There is a gap between training and test accuracies and they never cross. This is probably due to the difference in distribution between train and test data point. If these two sets are exactly similar we would probably get a very very narrow gap or even no gap. But in reality our train data and test data differs and so we get such gaps.