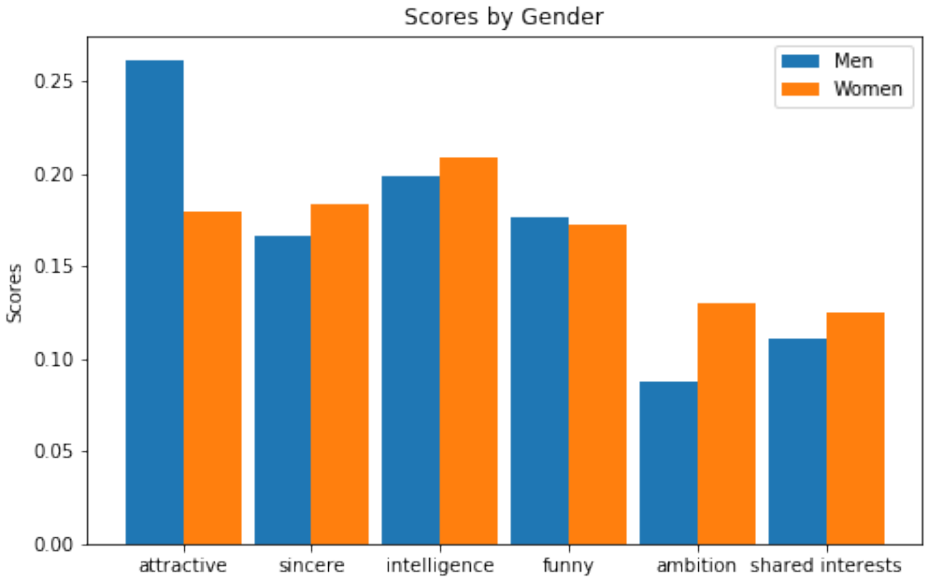
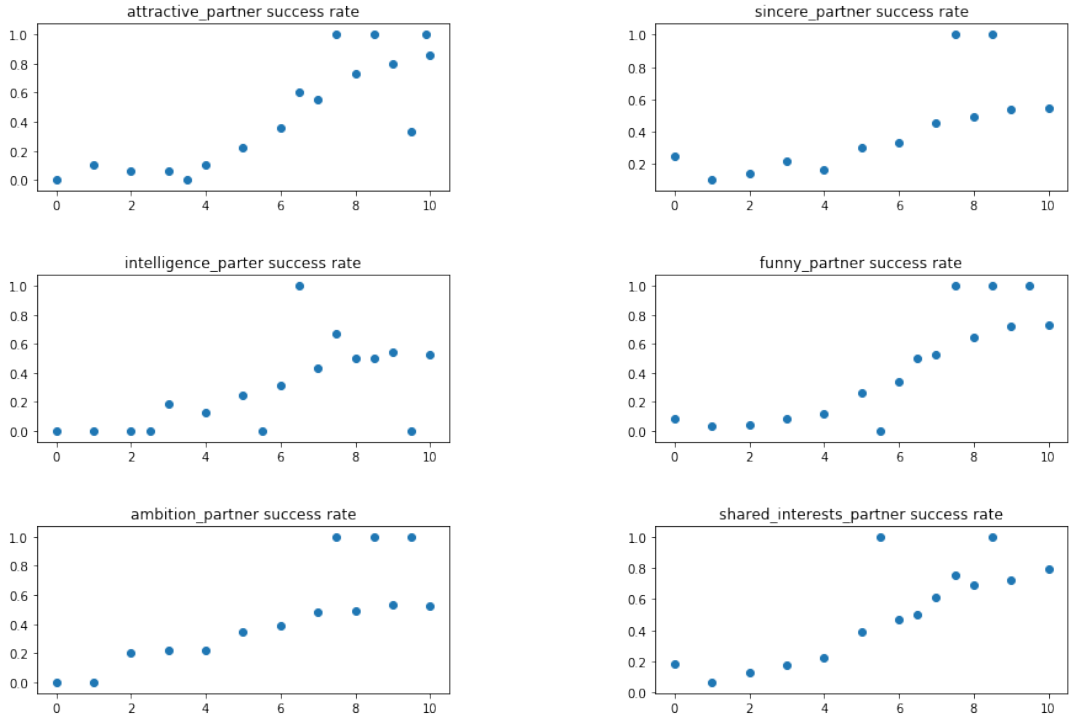
CS 573: Homework 2

Andrea Zapata - PUID : 0031827996

2 Visualizing interesting trends in data

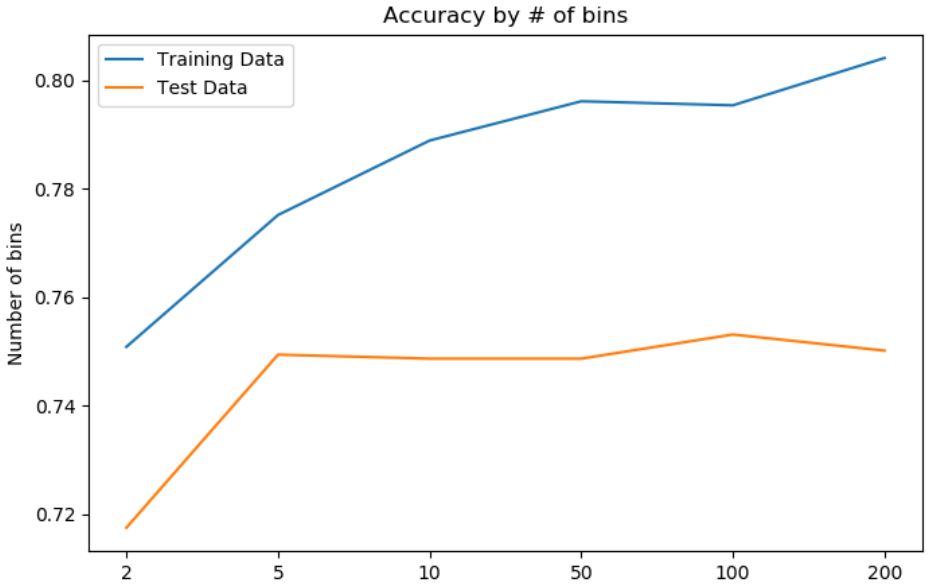


Overall, we can observe that the scores for men and women are very similar for the attributes sincere, intelligence, funny and shared interest. Men clearly favor attractiveness over all other attributes, women on the contrary prefer intelligence. The least important attribute for men is ambition, while for women is shared interests.



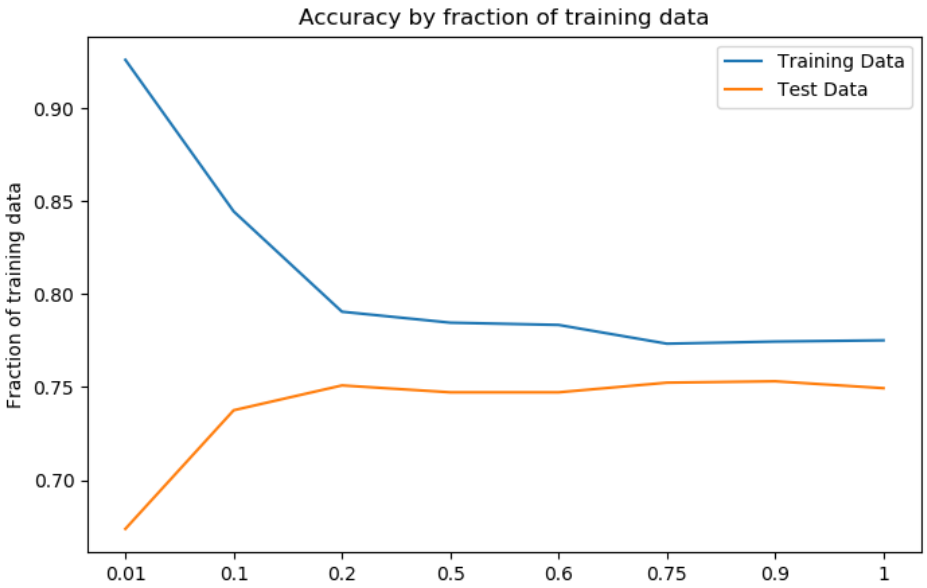
From the scatter plots, we can see that it exists a positive correlation between the score given for each of the attributes and the possibility to have a second date. At first sight, we could say that a high score (> 7) on attractive and funny partner gives the best chances for a second date, with a success rate over 0.8.

5 Implement a Naive Bayes Classifier



From the graph, we can observe that the number of bins influences the accuracy of the training data, which grows as the number of bins grows. As we increase the number of bins, our model becomes more

complex and will tend to learn the noise in our data. On testing data, we can see that after 5 bins, there is no significant grow on the accuracy. I think that the graph shows us that the optimal number of bins is 5 and increasing the number of bins will made us overfit the training data.



From this graph, we can see that the fraction of the training data used to train the Naive Bayes classifier influences the accuracy for both the training and test data. When we have a small training set, our model will likely achieve high accuracy for this set, but perform poorly on the test data. As our training data might not represent the underlying model of the whole dataset, our learned model is too naive. We also can see that the difference between the training and testing accuracy decreases as this fraction grows, because our model is becoming more complex and will be able to represent the underlying probabilities of our data.