DS 707
Classification Deliverable: Team 23
Dating Analytics

Rishabh Manoj (IMT2013035)
Revanuru Karthik (IMT2013033)
Nigel Fernandez (IMT2013027)
Anirudh Ravi (IMT2013005)

November 20, 2016

Contents

List of Figures

1 Classification on the Speed Dating Data Set

Our classification goal was defined as "Given a data point of two partners will they mutually accept each other". Since males and females could have different methodologies to accept partners in a date, we divided our model into two models based on gender, namely, (1) *male acceptance model* and (2) *female acceptance model*. Our analysis also considered whether race would be a factor in the acceptance process.

The *male acceptance model* would be trained on a chosen set of individual and common attributes of both the male and female partner involved in the date with the class label being whether the male participant accepted his female date partner. The *female acceptance model* was trained on the same attribute set with the exception being the class label changed to whether the female partner accepted her male date partner. These two models would be run for a new data point of two participants, if both the models return an acceptance value, the two participants would be mutually matched.

1.1 Choosing our Attributes

We chose our attributes based on our descriptive analytics results. This implied that we narrowed our attribute set to those attributes in a participant which were involved and highly correlated to the decision making process of the partner. Our attribute set consisted of three sets, namely, (1) how would you rate yourself on a scale of 1-10 on five attributes including attractiveness, fun, ambitiousness, sincerity and intelligence, (2) what do you think a partner expects on a scale of 1-10 on six attributes including attractiveness, fun, ambitiousness, sincerity, intelligence and shared interests, (3) rate yourself on a scale of 1-10 on 17 different interests including sports, movies, etc depending on your passion in them. Therefore our first attribute set contained $28 \cdot 2 = 56$ attributes and the class label which is binary (accepted or not).

We have a *samerace* attribute in our data set which is 1 if the participant prefers a partner of the same race or 0 otherwise. We included this attribute in our second choice of the attribute set to explore if race is a factor in the acceptance process. Our second attribute set contained $28 \cdot 2 = 56$ attributes, the *samerace* attribute and the class label which is binary (accepted or not).

2 Attribute Set excludes Race

2.1 SVM

Male Acceptance Model	
Sample Size	4040
Training Data	80%
Test Data	20%
Metric	Euclidean
Accuracy	69.230%

Female Acceptance Model	
Sample Size	4040
Training Data	80%
Test Data	20%
Metric	Euclidean
Accuracy	47.019%

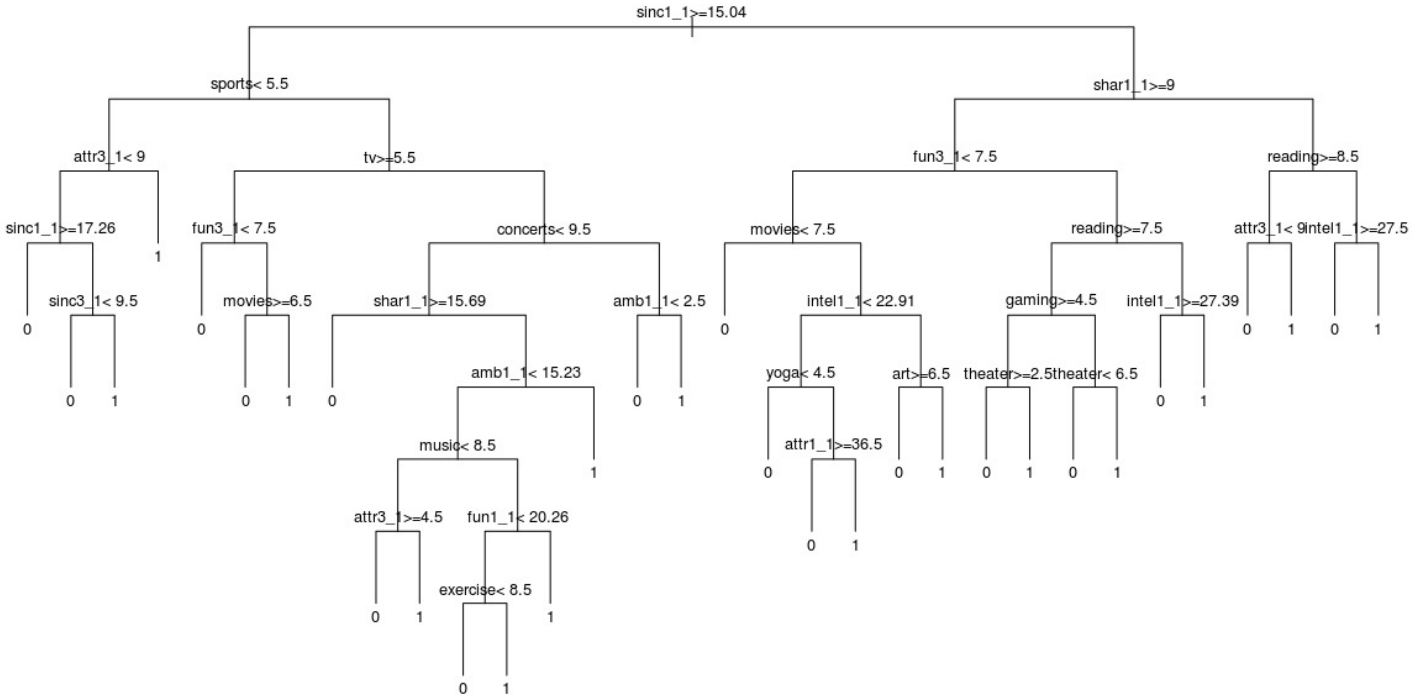
2.2 Naive Bayes

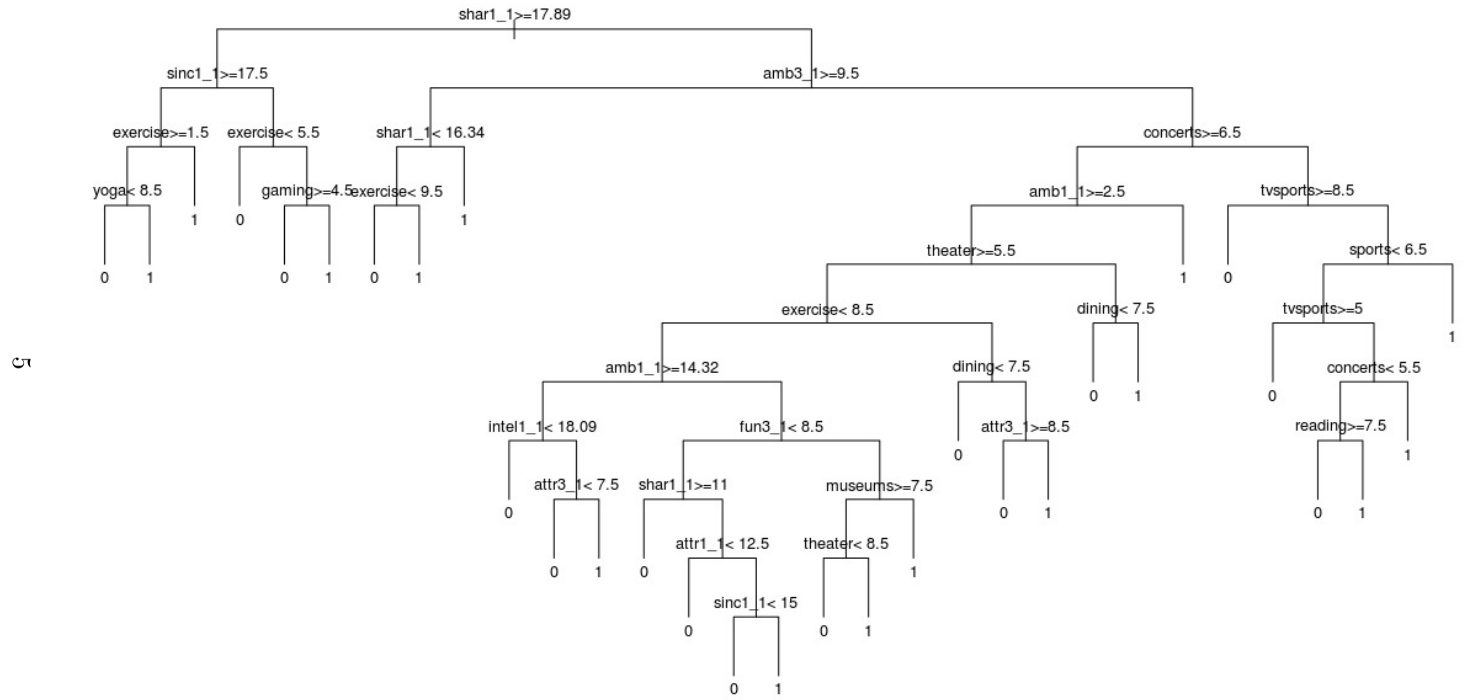
Male Acceptance Model	
Sample Size	4040
Training Data	80%
Test Data	20%
Metric	Euclidean
Accuracy	63.214%

Female Acceptance Model	
Sample Size	4040
Training Data	80%
Test Data	20%
Metric	Euclidean
Accuracy	62.981%

2.3 Decision Trees

Match for Males




Match for Females
shar1_1>=17.89
sinc1_1>=17.5
amb3_1>=9.5
exercise>=1.5
exercise< 5.5
shar1_1< 16.34
concerts>=6.5
yoga< 8.5
gaming>=4.5
exercise< 9.5
amb1_1>=2.5
tvsports>=8.5
theater>=5.5
sports< 6.5
exercise< 8.5
dining< 7.5
tvsports>=5
amb1_1>=14.32
dining< 7.5
concerts< 5.5
intel1_1< 18.09
fun3_1< 8.5
attr3_1>=8.5
reading>=7.5
attr3_1< 7.5
shar1_1>=11
museums>=7.5
attr1_1< 12.5
theater< 8.5
sinc1_1< 15

3 Attribute Set includes Race

3.1 SVM

Male Acceptance Model	
Sample Size	4040
Training Data	80%
Test Data	20%
Accuracy	62.857%

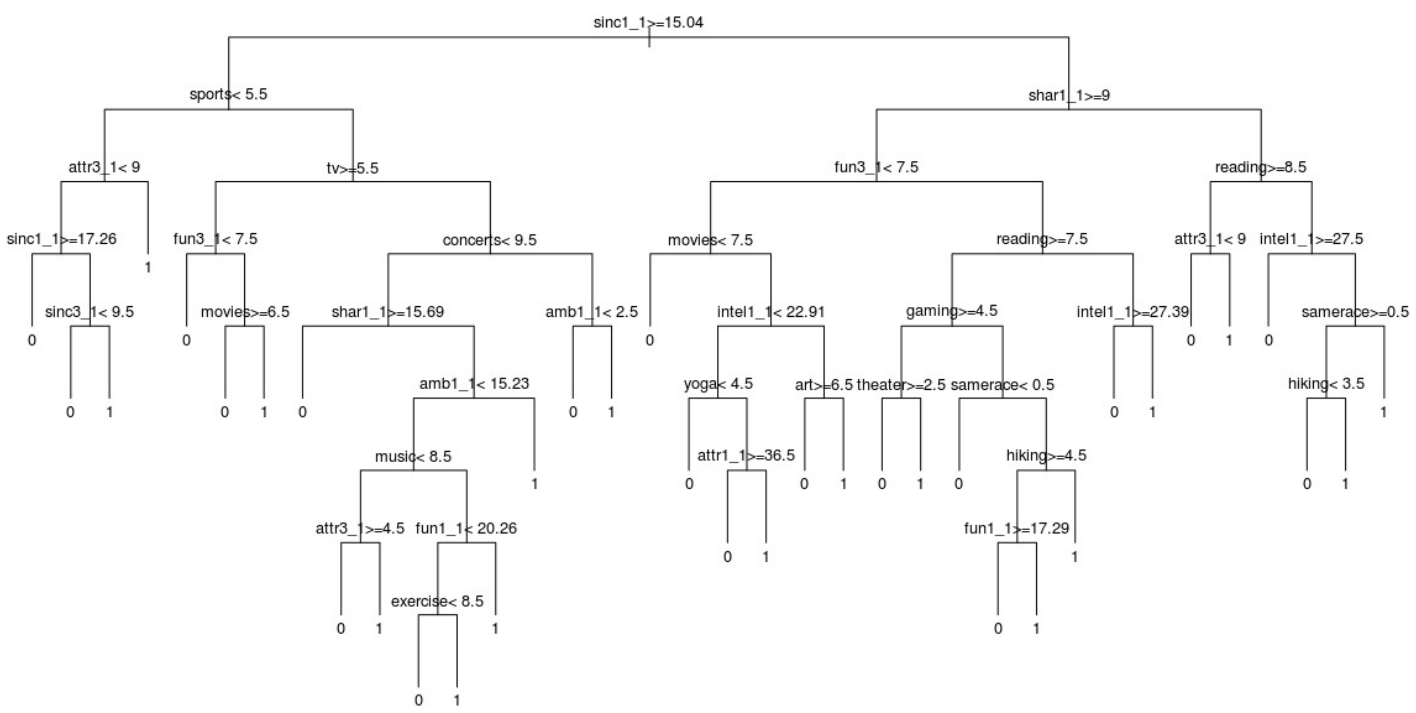
Female Acceptance Model	
Sample Size	4040
Training Data	80%
Test Data	20%
Accuracy	47.712%

3.2 Naive Bayes

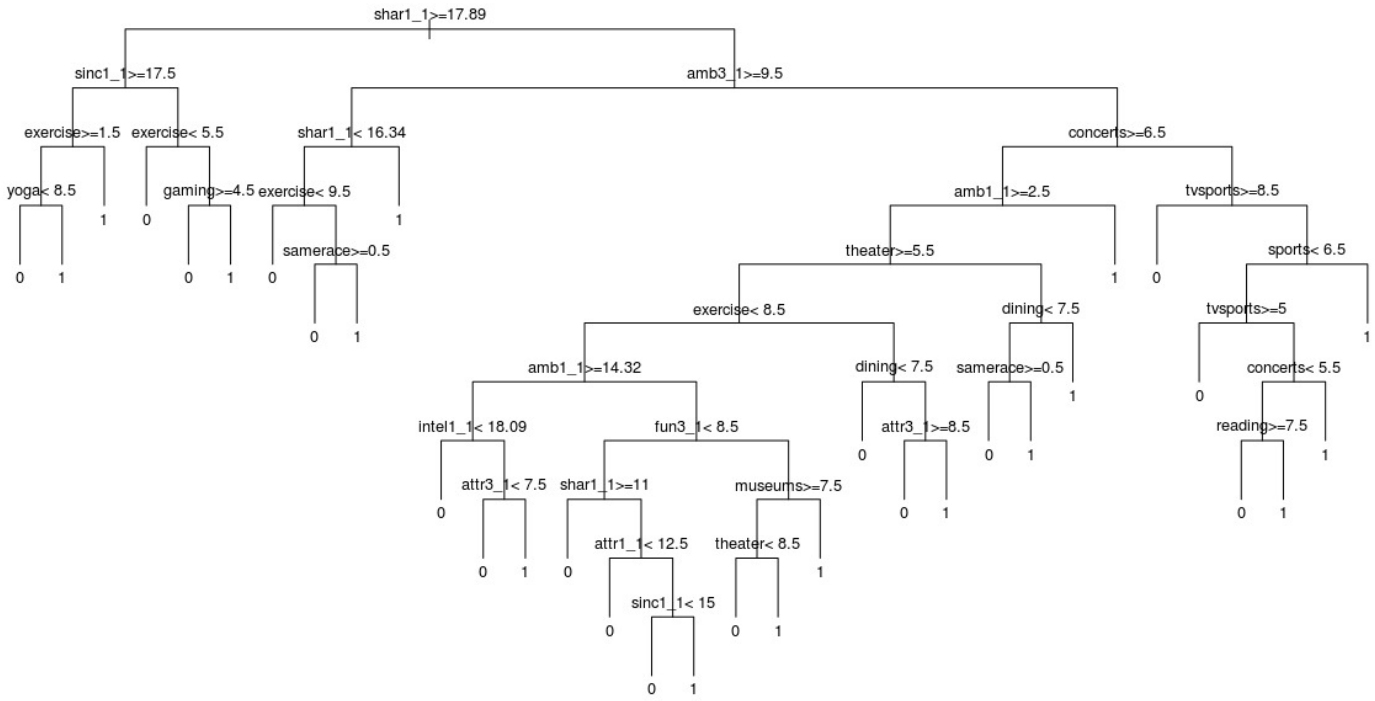
Male Acceptance Model	
Sample Size	4040
Training Data	80%
Test Data	20%
Accuracy	63.715%

Female Acceptance Model	
Sample Size	4040
Training Data	80%
Test Data	20%
Accuracy	63.287%

3.3 Decision Trees


Match for Males with Race
sinc1_1>=15.04
sports< 5.5
shar1_1>=9
attr3_1< 9
tv>=5.5
fun3_1< 7.5
reading>=8.5
sinc1_1>=17.26
fun3_1< 7.5
concerts< 9.5
movies< 7.5
reading>=7.5
attr3_1< 9
intel1_1>=27.5
sinc3_1< 9.5
movies>=6.5
shar1_1>=15.69
amb1_1< 2.5
intel1_1< 22.91
gaming>=4.5
intel1_1>=27.39
samerace>=0.5
amb1_1< 15.23
yoga< 4.5
art>=6.5
theater>=2.5
samerace< 0.5
hiking< 3.5
music< 8.5
attr1_1>=36.5
hiking>=4.5
attr3_1>=4.5
fun1_1< 20.26
fun1_1>=17.29
exercise< 8.5

Match for Females with Race



∞

4 PAC Learning

Theorem 4.1 *For any learning algorithm $\mathcal{A}$ with the hypothesis class $\mathcal{H}$ such that $|\mathcal{H}| = k$ for a fixed constant k, if $h^* = \mathcal{A}(D_n)$ is the output hypothesis on a training dataset D_n of size n then*

$$Pr(|R_e(h^*) - R(h^*)| > \epsilon) \leq 2k \cdot e^{-2\epsilon^2 n}$$

The above Theorem 4.1 implies:

- As the sample size of the training data set increases, the accuracy of our algorithm (learning algorithm) increases.
- As the training to test data ration increases, the accuracy of our algorithm (learning algorithm) increases.

We have shown these two results through two graphs 1 and 2 generated by simulation.

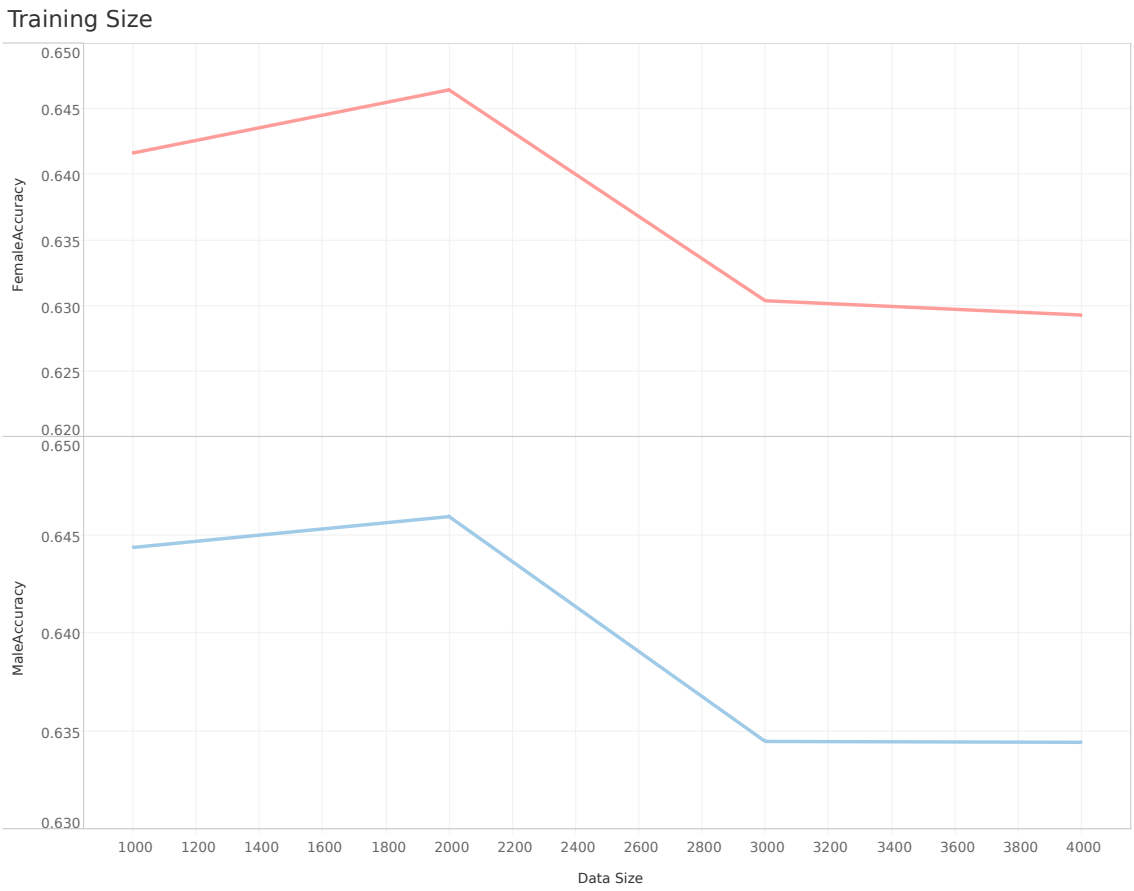


Figure 1: Effect of Sample Size on Accuracy

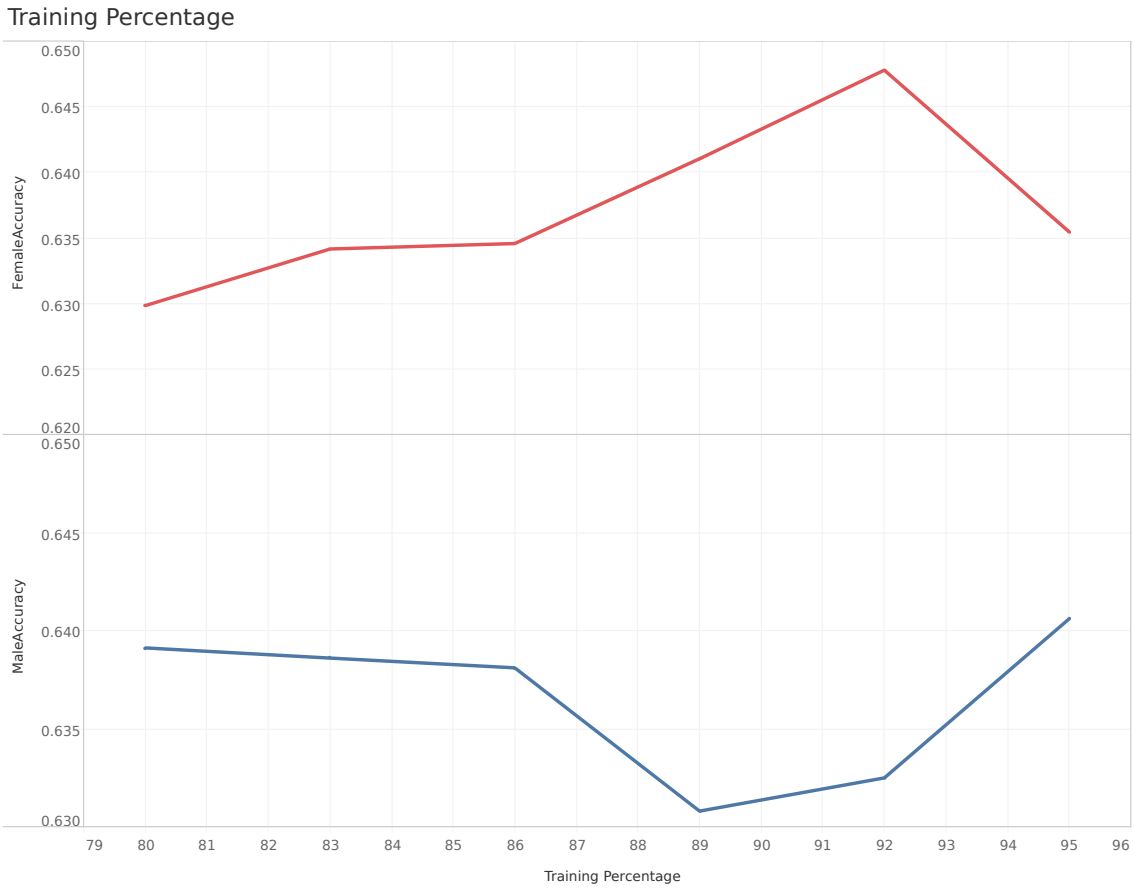


Figure 2: Effect of Training Set Percentage on Accuracy

5 Conclusion

As observed in our classification experiment, our *male acceptance model's* accuracy and our *female acceptance model's* accuracy remained similar when the *samerace* attribute was included as part of the feature set. This might be because of the minimal correlation between racial preference and the decision of accepting a partner thus indicating that race does not play a role in the dating process.