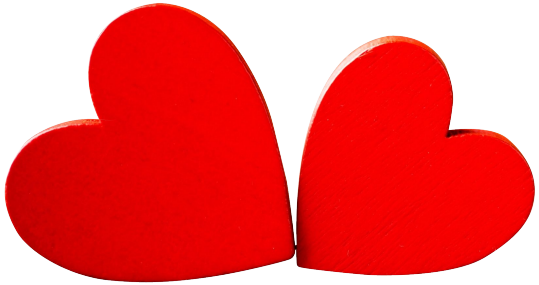
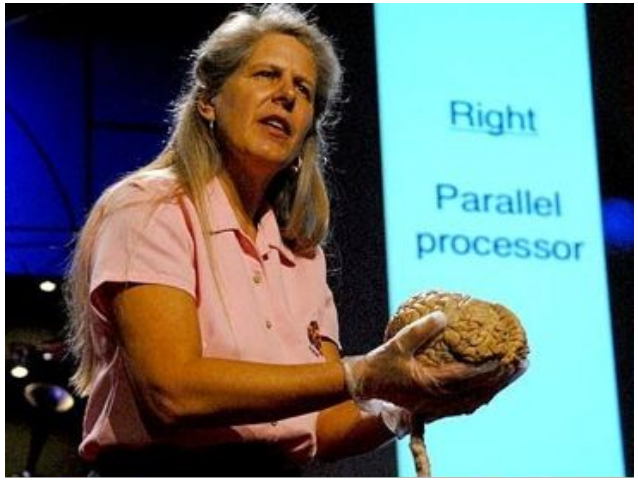
Uncovering the Truth of Love: A Modern Approach

By Kyle Wang, Feiyu Yue, Tz-Ruei Liu

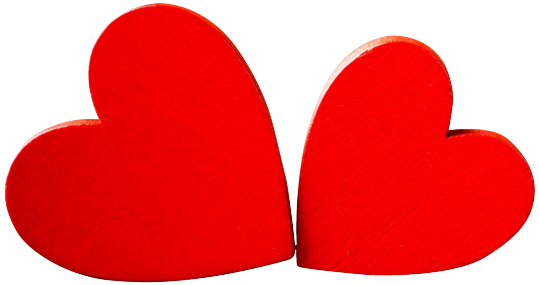
Motivation



Source: https://pngriver.com/download-heart-love-png-background-image-71571/



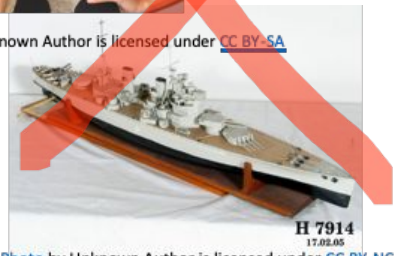
Motivation



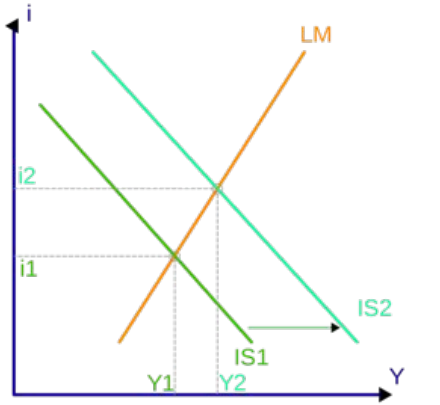
Source: https://pngriver.com/download-heart-love-png-background-image-71571/

This Photo by Unknown Author is licensed under CC BY-NC-ND

This Photo by Unknown Author is licensed under CC BY-SA



This Photo by Unknown Author is licensed under CC BY-NC-ND



This Photo by Unknown Author is licensed under CC BY-SA

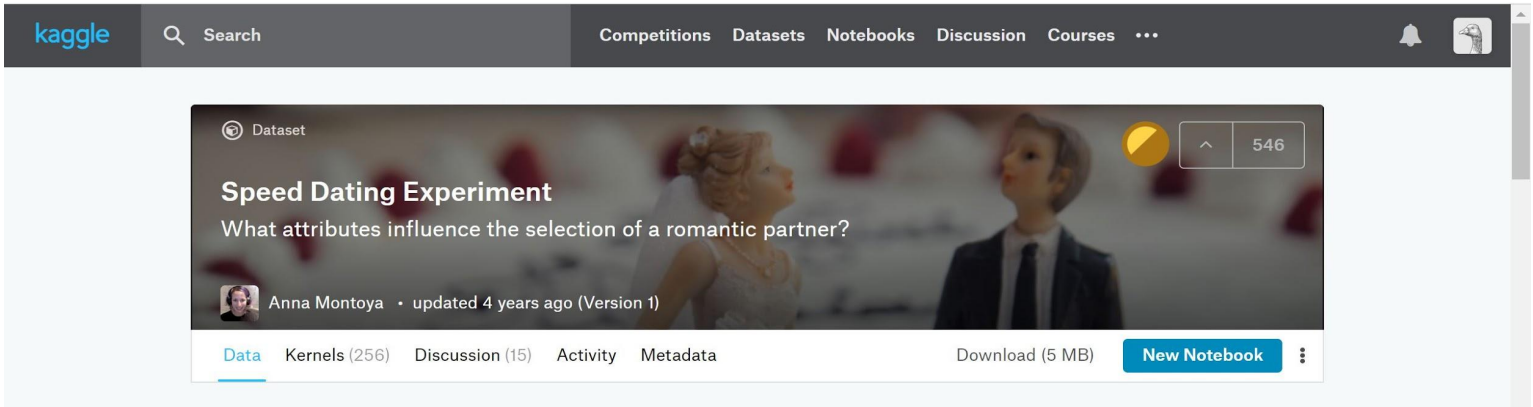
Goal

~~To test if Machines are better at finding soul mates than ourselves~~

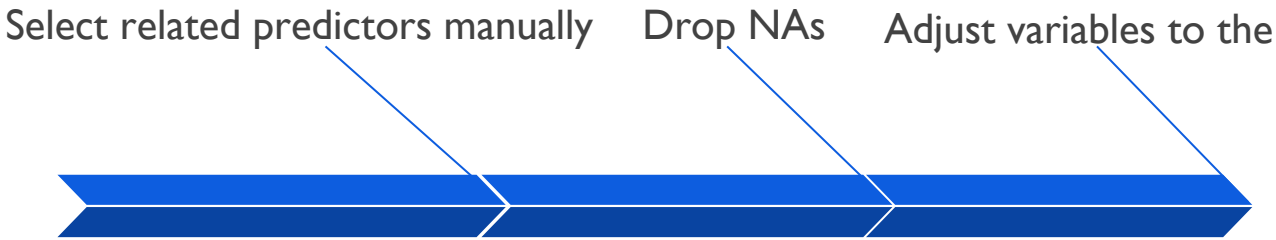
- To use models to predict whether two people would match

Overview of Dataset

- Our dataset is provided by Kaggle dataset: Speed Dating Experiment. (https://www.kaggle.com/annavictoria/speed-dating-experiment)
- It was first compiled by professors Ray Fisman and Sheena Iyengar from Columbia for their paper Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment.



Data Cleaning

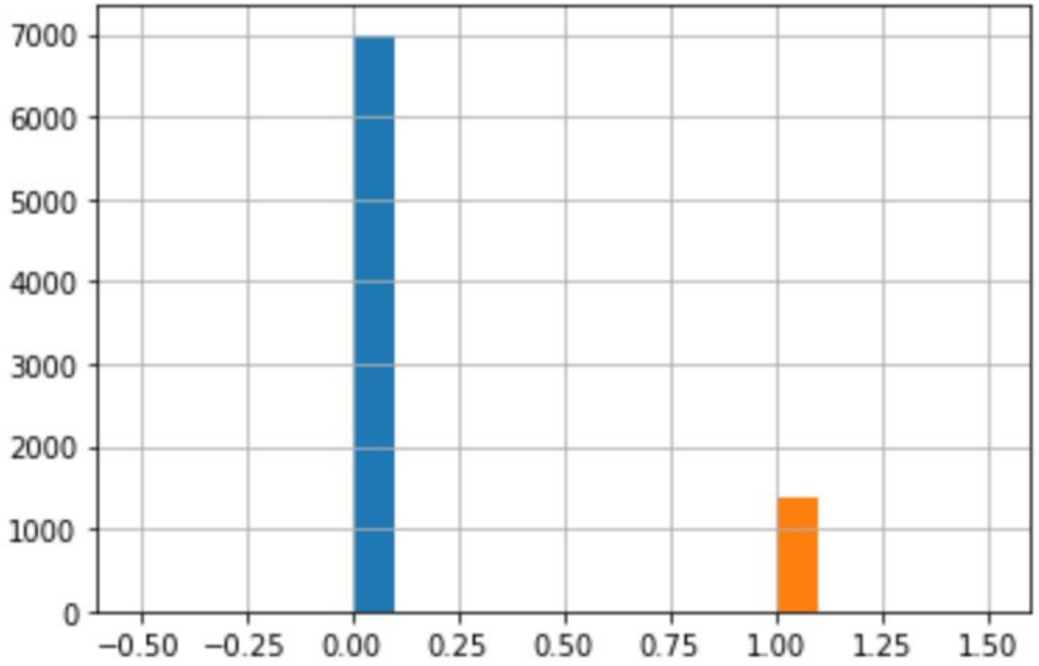



```
idnum = ['iid','pid']
attributes = ['gender', 'age', 'field', 'race',
attitudes = ['goal','exphappy']
label = ['match']

date = df_full[idnum+attributes+attitudes+label]
date.head()
print(date.shape)

# After dropna()
date = date.dropna().drop_duplicates()
print(date.shape)
```

```
(8378, 43)
(8124, 43)
```

- Each row includes 40 predictors, including age, field of study, race, etc., and also the person's id that he/she dated with. (the rest 3 are 'id', 'partner id', and 'match')

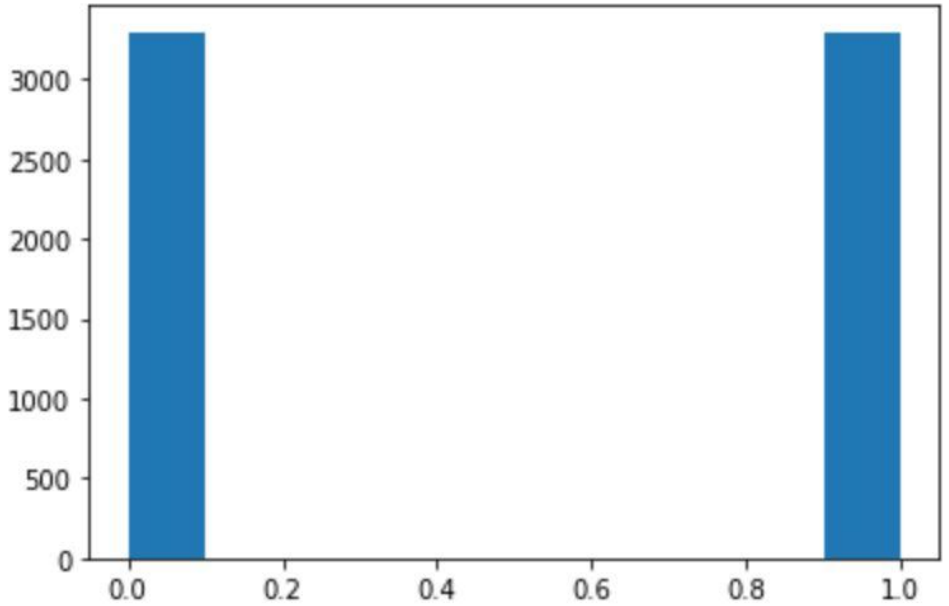
20%



Sadly, only 20% of speed dating partners match.

Finding true love is hard :(

50%



Our response variable is thus very imbalanced.

We apply imbalanced-learn over our dataset to offset this imbalance.

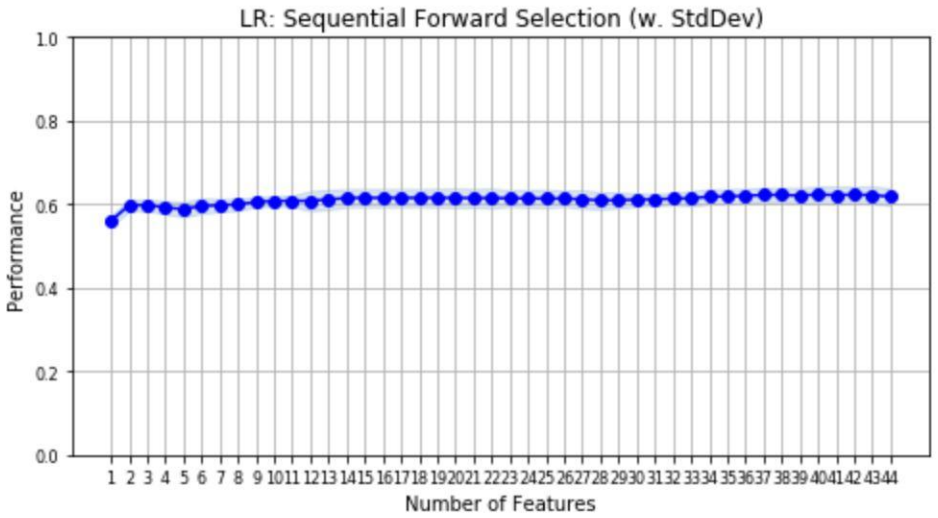
Two Approaches

The following models would all contain two approaches:

1. The difference in features as predictors (denote as **diff**)
2. Both partners' features as predictors (denote as **concat**)

For both approaches, the response variable is whether the two partners **match**. (Yes/No)

Model 1: Logistic Regression - Diff

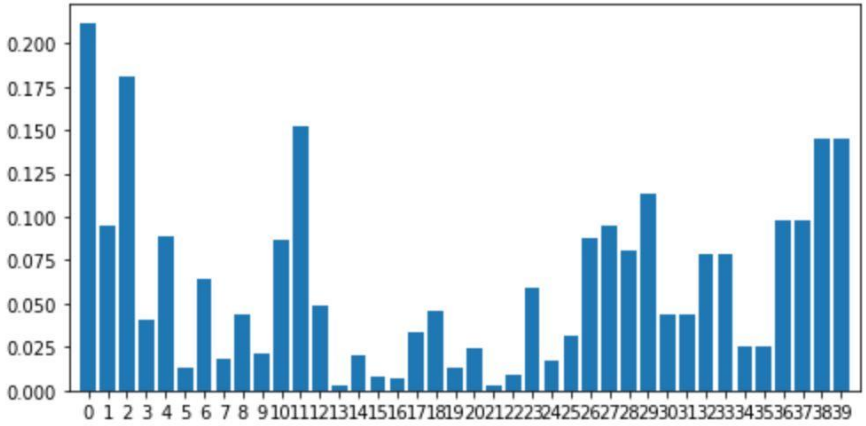


- Apply forward sequential search.
- Check how many features to select that give the best score.
- Test accuracy remains about 60% no matter how many features one select:(

```
Validation Accuracy 62.84%
Test Accuracy 63.00%
```

All predictors don’t contribute too much

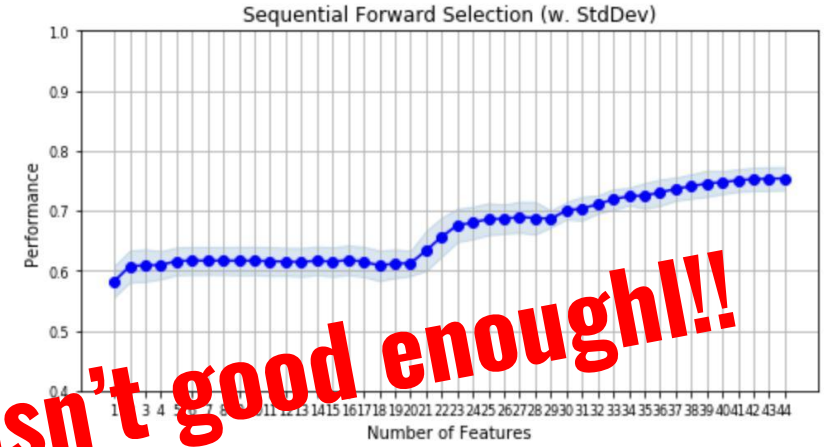
The only predictor that’s greater than 0.2 is ageDiff. People with similar ages are more likely to match, but still no significant at all...



Absolute values of coefficients of all predictors

Model 1: Logistic Regression - Concat

- Again apply forward sequential search.
- Seems that applying more predictors would give us better accuracy.
- Test accuracy improves compared to [illegible] still.



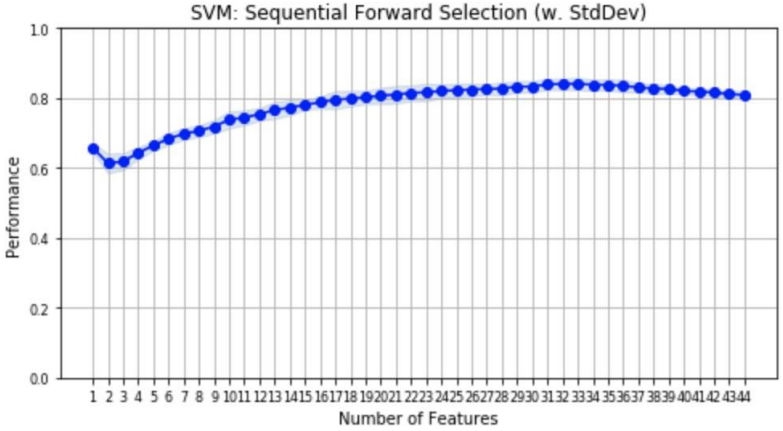
Logistic Regression isn't good enough!!

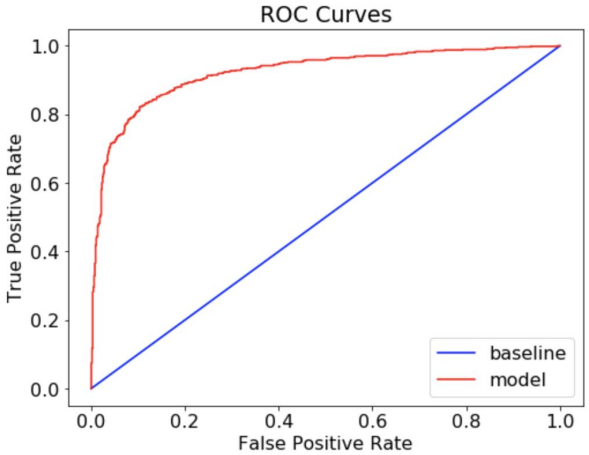
Validation Accuracy 74.31%
Test Accuracy 74.53%

Model 2: SVM - Diff

- Apply sequential forward selection(SFS) on the normalized training data
- k = 33 gives the best score

Training Accuracy 92.01%
Validation Accuracy 83.98%
Test Accuracy 84.61%





Race? From? Career? --Who Cares.

It seems that people care more about personalities instead of origins...

Dropped features

0	1	2	3	4
career_cDiff_n	career_cDiff_y	exerciseDiff	field_cDiff_n	field_cDiff_y

5	6	7	8	9	10
fromDiff_n	fromDiff_y	goalDiff_n	goalDiff_y	raceDiff_n	raceDiff_y

Selected features

0	ageDiff	9	artDiff	18	musicDiff	27	attr3_1Diff
1	impraceDiff	10	hikingDiff	19	shoppingDiff	28	sinc3_1Diff
2	impreligDiff	11	gamingDiff	20	yogaDiff	29	intel3_1Diff
3	dateDiff	12	clubbingDiff	21	attr1_1Diff	30	fun3_1Diff
4	go_outDiff	13	readingDiff	22	sinc1_1Diff	31	amb3_1Diff
5	sportsDiff	14	tvDiff	23	intel1_1Diff	32	exphappyDiff
6	tvsportsDiff	15	theaterDiff	24	fun1_1Diff		
7	diningDiff	16	moviesDiff	25	amb1_1Diff		
8	museumsDiff	17	concertsDiff	26	shar1_1Diff		

Model 2: SVM - Concat

- Apply model selection but without feature selection

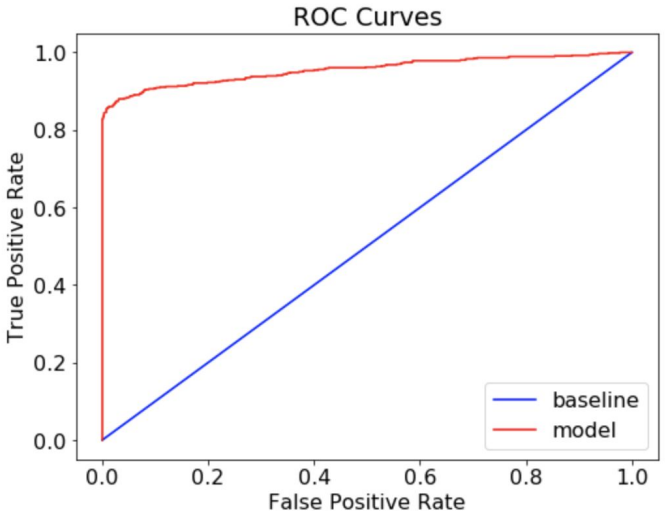
```
from sklearn.model_selection import GridSearchCV
clf = svm.SVC()
parameters = {'kernel':('linear', 'rbf'), 'C':[0.5, 1]}
gs_svm = GridSearchCV(clf, parameters, cv = 5, error_score=0,
                      n_jobs=-1, verbose = 10)
```

```
print(gs_svm.best_params_)
```

```
{'C': 1, 'kernel': 'rbf'}
```

```
Training accuracy: 94.59%
Test accuracy: 92.19%
```

Better than Diff !



Model 3: Random Forest

First try: Overfit

```
random_classifier = RandomForestClassifier()

random_classifier.fit(X_train, y_train)
```

```
Transf. training accuracy: 99.29%
Transf. test accuracy: 86.58%
```

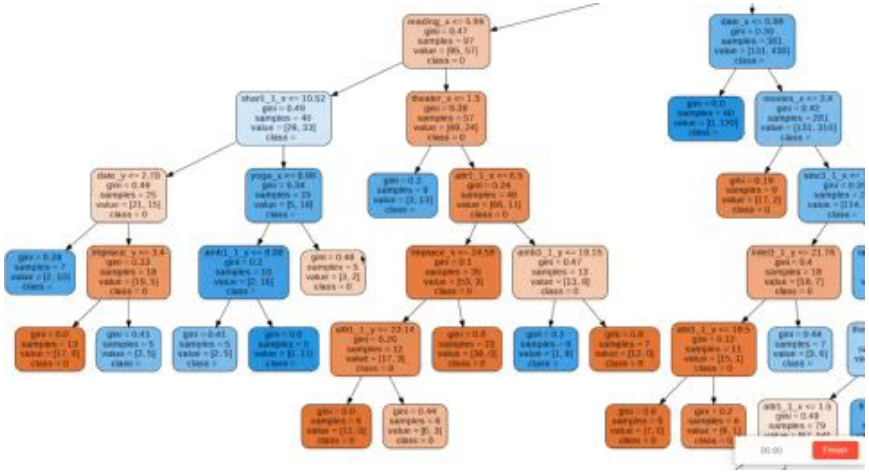
Then: GridSearch with cv=5

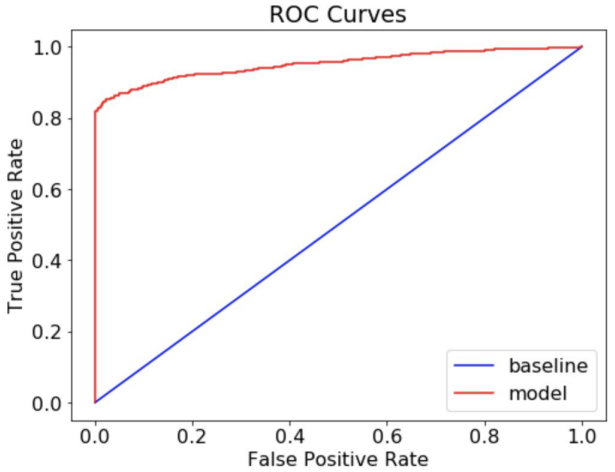
```
parameters = { 'max_features':np.arange(35,45),
               'n_estimators':[100],
               'min_samples_leaf': [5,10,15,20,25]}

rf = GridSearchCV(random_classifier, parameters,
                  cv = 5, error_score=0,
                  n_jobs=-1, verbose = 10)
```

Model 3: Random Forest - Diff

```
Transf. training accuracy: 97.75%
Transf. test accuracy: 90.83%
{'max_features': 35, 'min_samples_leaf': 5, 'n_estimators': 100}
```

```
modelrf = RandomForestClassifier(max_features=36, min_samples_leaf= 5, n_estimators=100)
modelrf.fit(X_train, y_train)
```

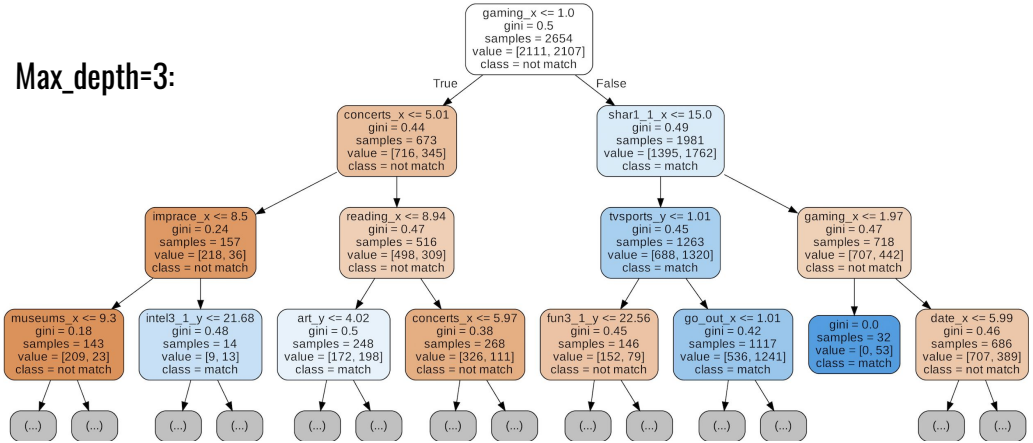


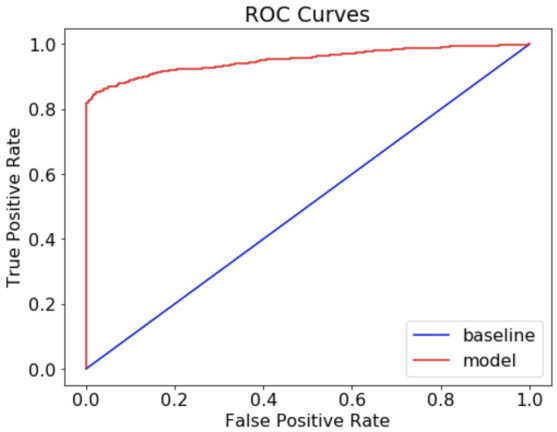


Model 3: Random Forest - Concat

```
Transf. training accuracy: 98.70%
Transf. test accuracy: 91.43%
{'max_features': 154, 'min_samples_leaf': 5, 'n_estimators': 600}
```

Max_depth=3:





Compared to the model generated based on differences, the accuracy is slightly higher. The ROC curve looks almost the same.

Conclusion

- Logistic Regression is okay, while SVM and Random Forest look good!
- It is not enough to just focus on the similarity between two participants.
- Perform well on this dataset, but probably not in reality.
- Love seems to be predictable using our models (?)

Future Improvements

1. Random Forest & SVM: Try dimensionality reduction methods to reduce overfitting.

2. We will try XGBoost to see how it performs.

3. For the concat approach, test it on a pair that neither of them present in the training set.

Q & A