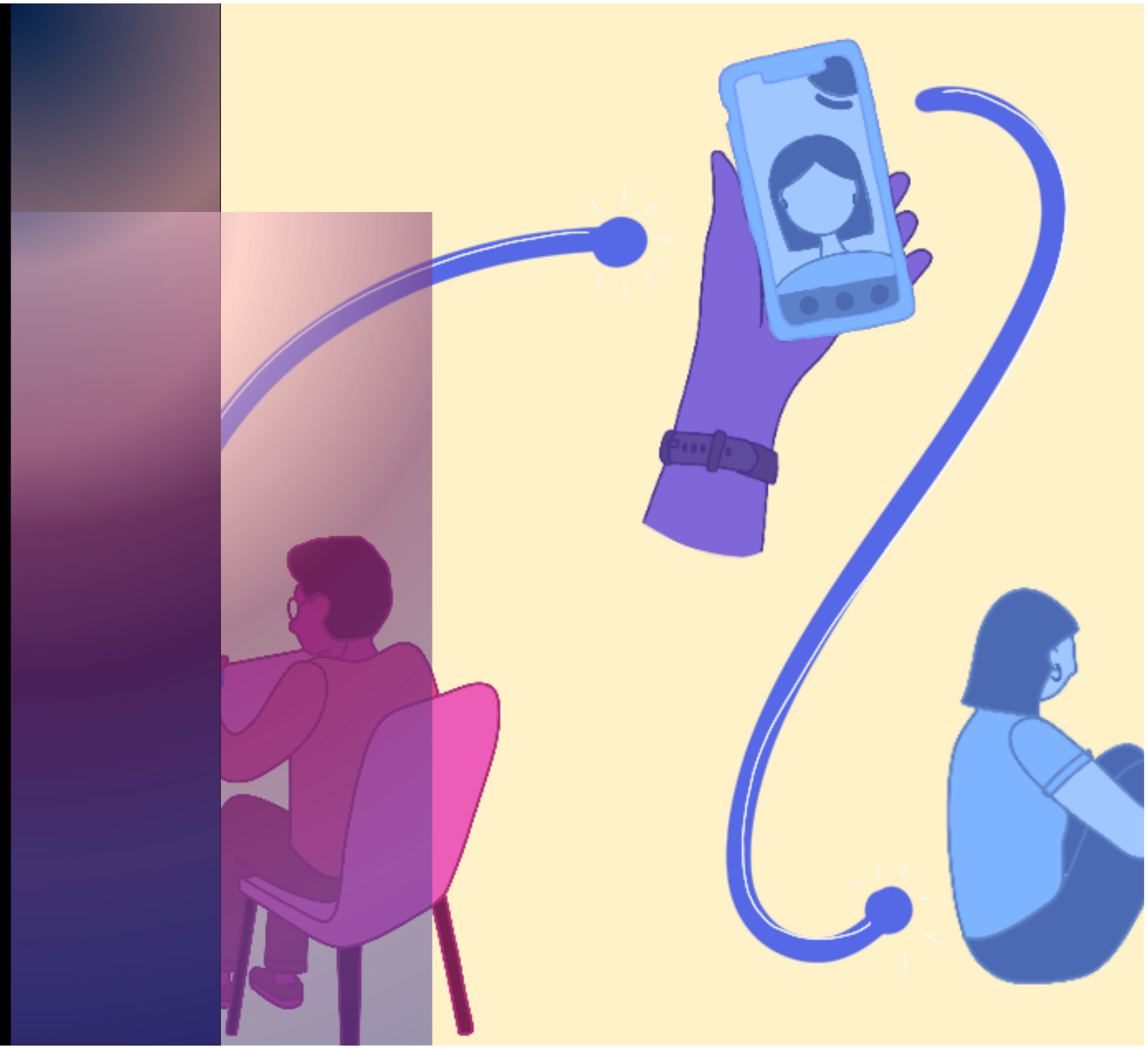
Speed Dating Predictions

A prediction model with an analysis of what drives people to choose a mate.



Business Idea : Speed Dating Match Prediction.

Importance : We find this as an interesting topic since there are already serval dating apps that show potential matches based on individual needs, likes and dislikes. It will also give us insight into human behaviour and understand what factors are important to people when it comes to choosing potential mates.

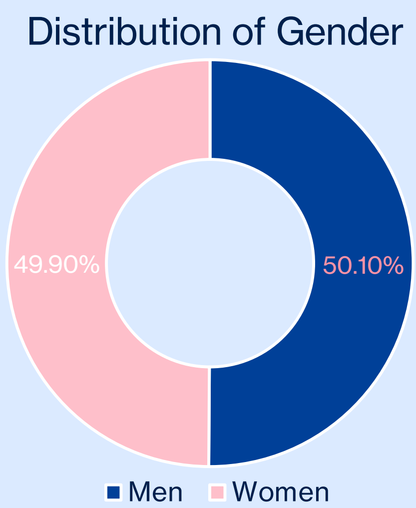
Data Resources: For this project we are using a dataset that was compiled by Columbia Business School professors Ray Fisman and Sheena Iyengar for their paper Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment.

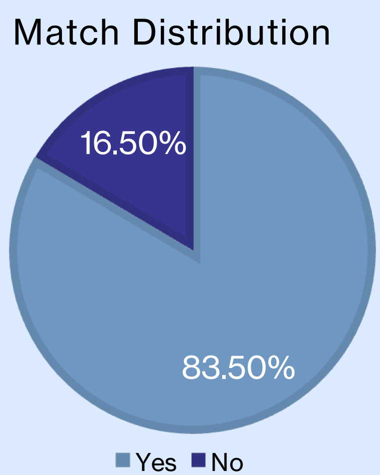
As part of their study the subjects were asked to fill a survey when they signed up for the Speed Dating Event and fill a scorecard on various attributes on each of their potential matches. At the end of the night subjects were matched with potential partners if both persons said yes. This experiment was conducted on various days with different groups of people and collated into one dataset.

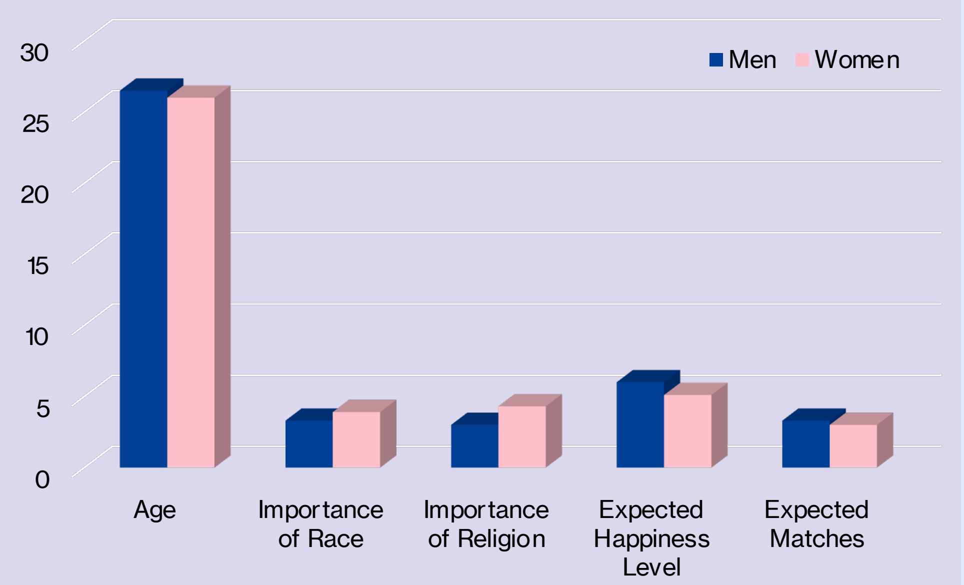
Our goal is to predict the result of matching as "Yes" or "No"

Some Information About our Data

- 84 attributes and 1 target(match) column
- 76 columns with null values and 7 object type columns.
- We will eventually drop columns with more than 20% missing values and fill the remaining null values with the mean of the column and create dummy variables using one hot encoding in Data Processing Part I.

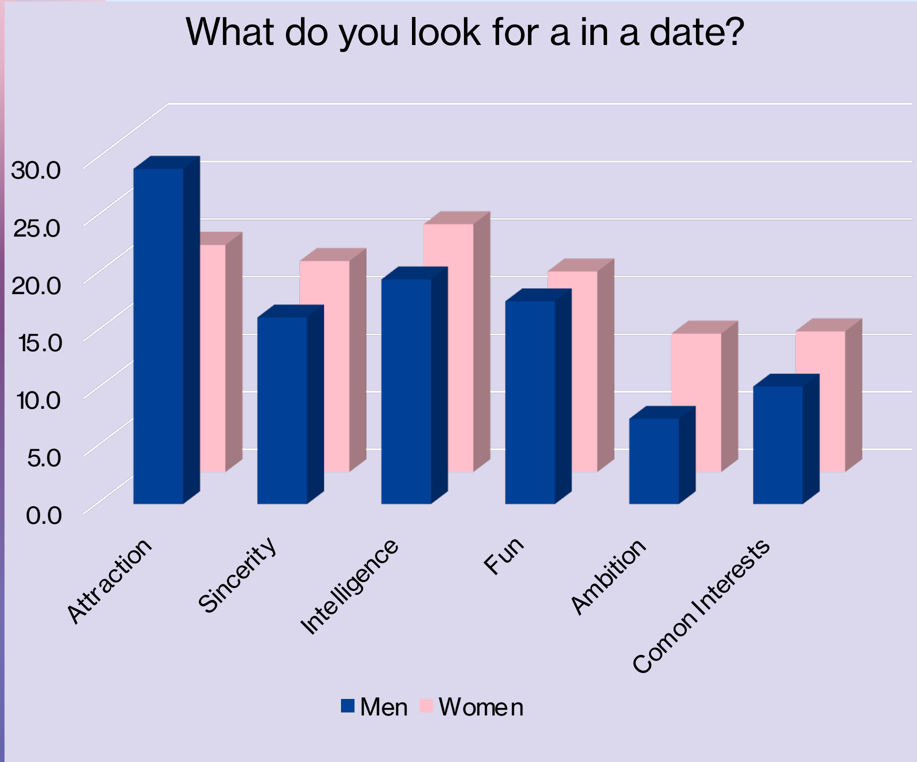


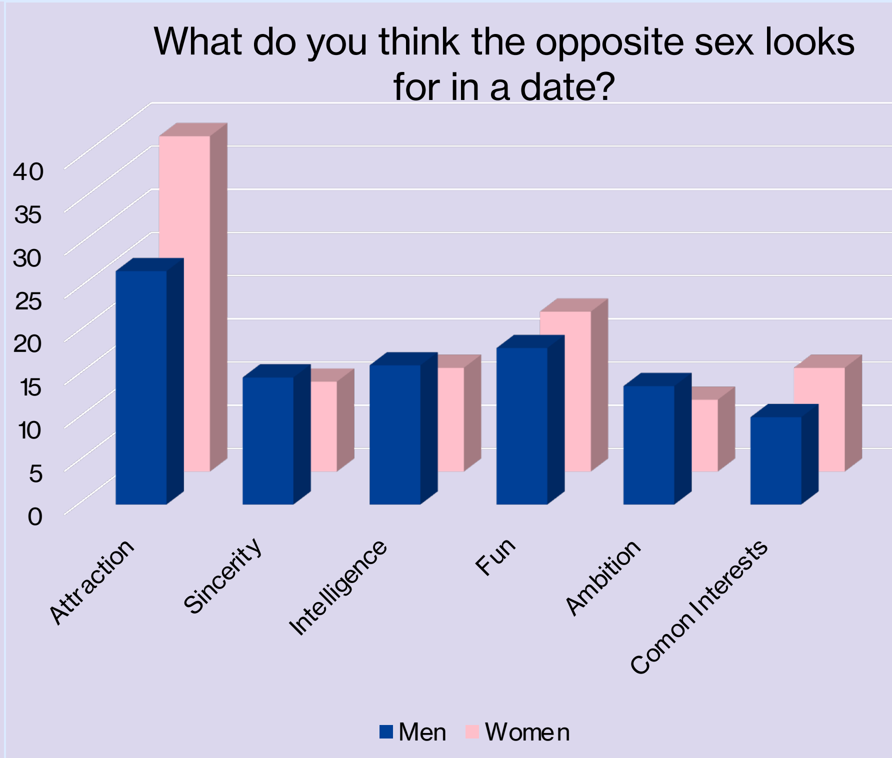


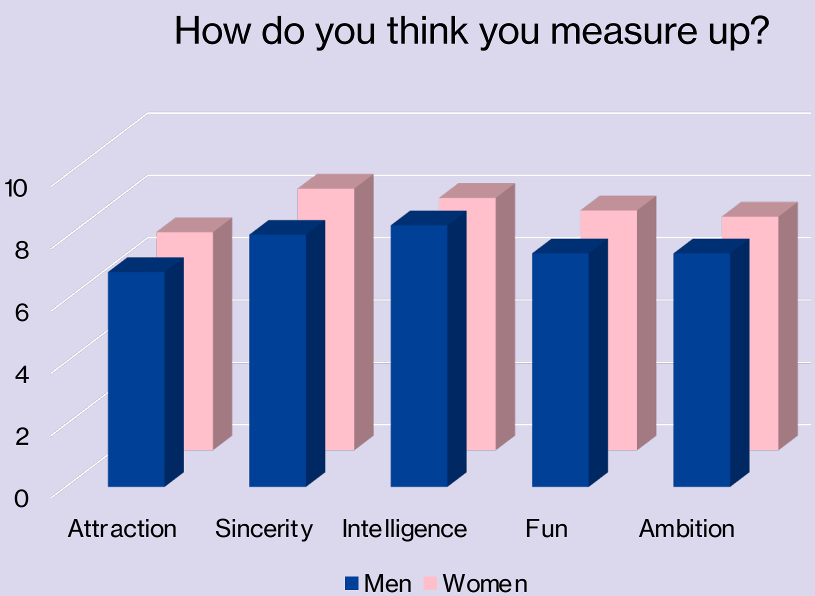


Gender Differences in Views & Perception

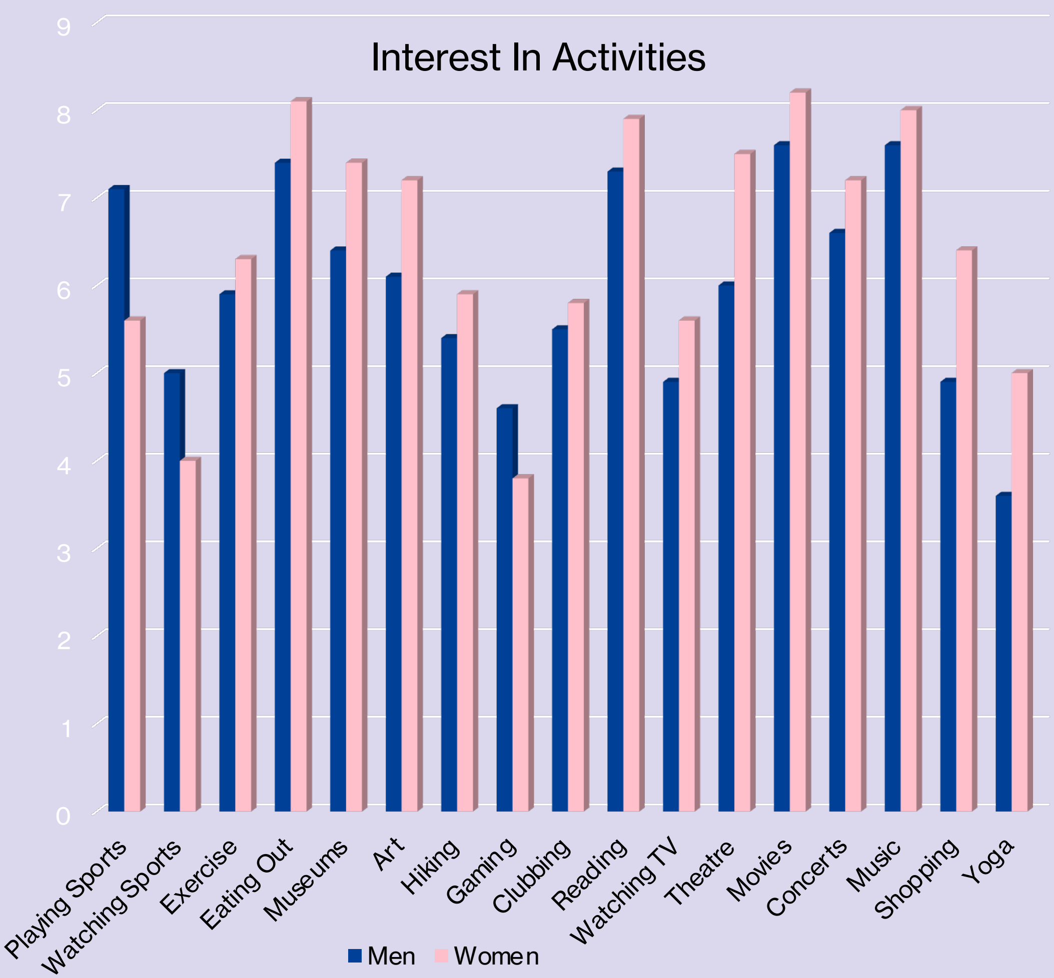
The Participants were asked to distribute 100 points across the below attributes in response to each question.







The participants view of their own attributes.



All this information was collected from the participants as part of the signing up process before the event.

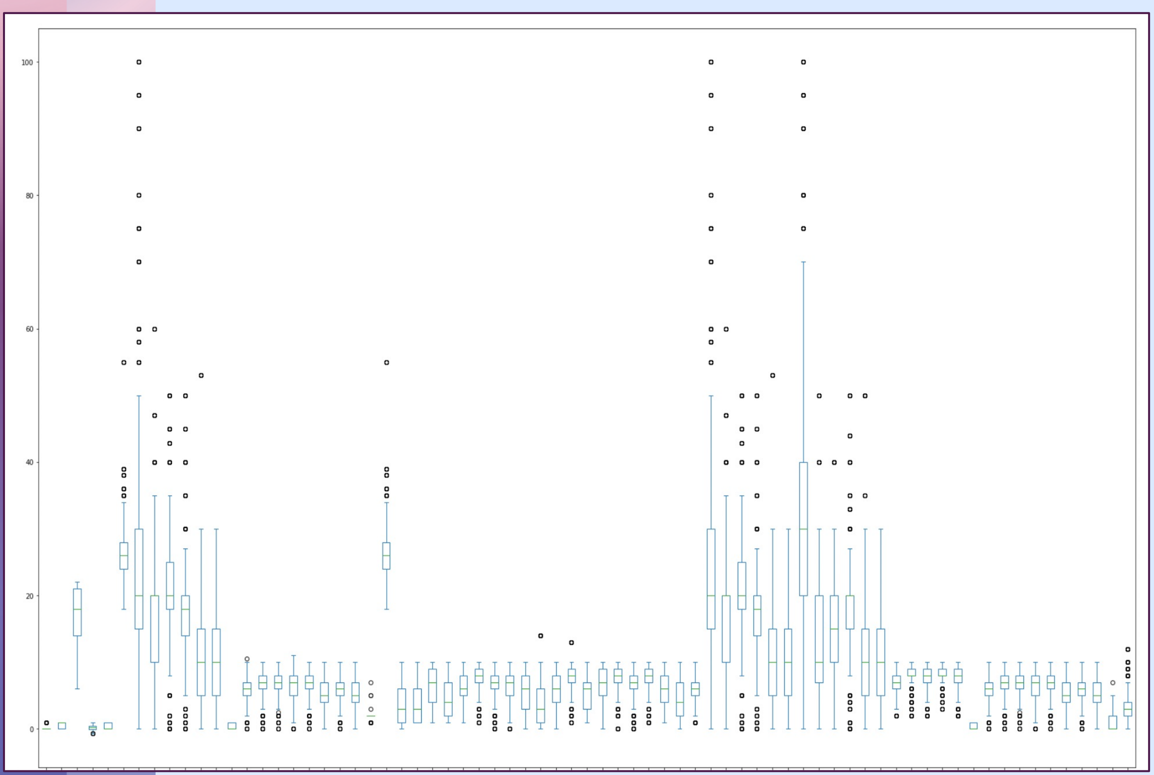
The participants chose to match or not with the person they met at the end of their 4-minute speed date.

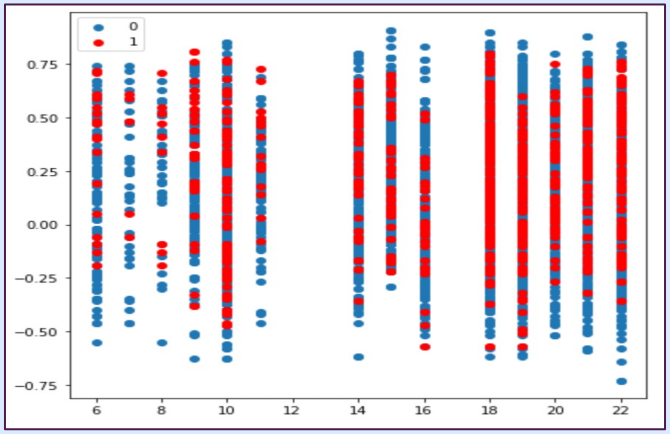
Data Processing Part II

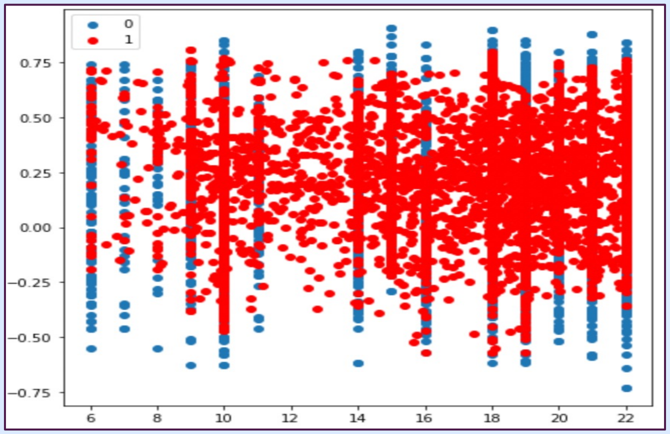
Having explored our data we then checked for any outliers that could cause our model to be overfit and/or underperform.

After Dealing with the outliers using Isolation Forest we began the process of fitting our models and tuning them.

We have also resampled our data using Synthetic Minority Oversampling Technique (SMOTE)







Principal Component Analysis Results

Logistic Regression

```
Classification report:
              precision    recall  f1-score   support

           0       0.83      0.69      0.76      1722
           1       0.13      0.24      0.17       323

    accuracy                           0.62      2045
   macro avg       0.48      0.47      0.46      2045
weighted avg       0.72      0.62      0.66      2045

Confusion matrix:
 [[1194  528]
 [ 244   79]]
```

```
Classification report:
              precision    recall  f1-score   support

           0       0.83      1.00      0.91      1524
           1       0.00      0.00      0.00       306

    accuracy                           0.83      1830
   macro avg       0.42      0.50      0.45      1830
weighted avg       0.69      0.83      0.76      1830

Confusion matrix:
 [[1524    0]
 [ 306    0]]
```

Decision Tree Classifier

```
Classification report:
              precision    recall  f1-score   support

           0       0.83      0.76      0.79      1703
           1       0.16      0.23      0.19       342

    accuracy                           0.67      2045
   macro avg       0.50      0.49      0.49      2045
weighted avg       0.72      0.67      0.69      2045

Confusion matrix:
 [[1289  414]
 [ 263   79]]
```

```
Classification report:
              precision    recall  f1-score   support

           0       0.83      0.81      0.82      1532
           1       0.15      0.17      0.16       298

    accuracy                           0.71      1830
   macro avg       0.49      0.49      0.49      1830
weighted avg       0.72      0.71      0.71      1830

Confusion matrix:
 [[1240  292]
 [ 246   52]]
```

Naïve Bayes Classifier

```
Classification report:
              precision    recall  f1-score   support

           0       0.83      1.00      0.91      1695
           1       0.30      0.01      0.02       350

    accuracy                           0.83      2045
   macro avg       0.56      0.50      0.46      2045
weighted avg       0.74      0.83      0.75      2045

Confusion matrix:
 [[1688    7]
 [ 347    3]]
```

```
Classification report:
              precision    recall  f1-score   support

           0       0.85      1.00      0.92      1546
           1       0.25      0.00      0.01       284

    accuracy                           0.84      1830
   macro avg       0.55      0.50      0.46      1830
weighted avg       0.75      0.84      0.77      1830

Confusion matrix:
 [[1543    3]
 [ 283    1]]
```

KNN Classifier

```
Classification report:
              precision    recall  f1-score   support

           0       0.83      0.99      0.90      1683
           1       0.38      0.03      0.06       362

    accuracy                           0.82      2045
   macro avg       0.60      0.51      0.48      2045
weighted avg       0.75      0.82      0.75      2045

Confusion matrix:
 [[1665   18]
 [ 351   11]]
```

```
Classification report:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91      1538
           1       0.00      0.00      0.00       292

    accuracy                           0.84      1830
   macro avg       0.42      0.50      0.46      1830
weighted avg       0.71      0.84      0.77      1830

Confusion matrix:
 [[1537    1]
 [ 292    0]]
```

Our Process

We used four machine learning models to classify our data into match and no match. We ran each of our models using four versions of our data to get the best model and accuracy.

Logistic Regression

Decision Trees Classifier

Naïve Bayes Classifier

K Nearest Neighbors

1. **Benchmark Data :** This data is the original dataset in which we have not dropped any columns or rows.

2. **Cleaned & Processed Data :** In this set we dropped columns and rows with logically and those with large volumes of null values(20% & above) and treated our dataset for outliers.

3. **Resampled Data :** We resampled our dataset using Synthetic Minority Oversampling Technique (SMOTE)

4. **Data using Extra Trees Classifier as a feature selection method :** Using this method we got the top 12 attributes which we then used for our machine learning models.

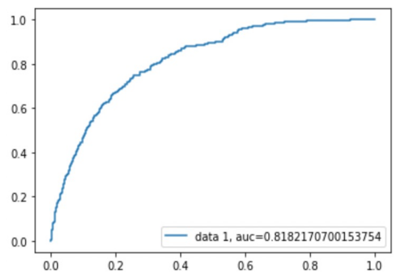
Logistic Regression Results

BENCHMARK RESULTS

Classification report:

	precision	recall	f1-score	support
0	0.86	0.96	0.91	1697
1	0.57	0.24	0.33	348
accuracy			0.84	2045
macro avg	0.72	0.60	0.62	2045
weighted avg	0.81	0.84	0.81	2045

Confusion matrix:
```
 [[1636   61]
 [ 266   82]]
```
Accuracy Score:
0.840
AUC:
0.8182170700153754

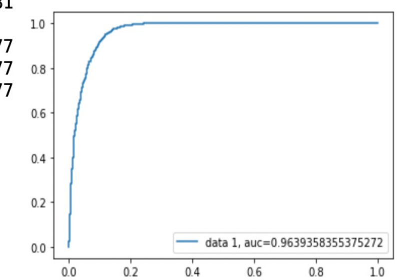


RESAMPLED DATA

Classification report:

	precision	recall	f1-score	support
0	0.96	0.87	0.91	1596
1	0.87	0.96	0.91	1481
accuracy			0.91	3077
macro avg	0.92	0.91	0.91	3077
weighted avg	0.92	0.91	0.91	3077

Confusion matrix:
```
 [[1384  212]
 [  58 1423]]
```
Accuracy Score:
0.912
AUC:
0.9639358355375272

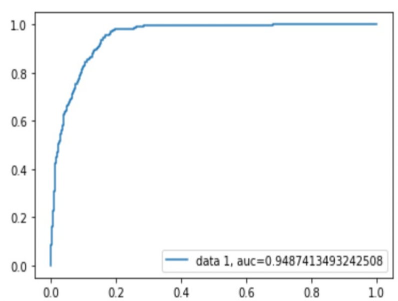


CLEANED PROCESSED DATA

Classification report:

	precision	recall	f1-score	support
0	0.94	0.95	0.94	1544
1	0.70	0.66	0.68	286
accuracy			0.90	1830
macro avg	0.82	0.81	0.81	1830
weighted avg	0.90	0.90	0.90	1830

Confusion matrix:
```
 [[1462   82]
 [  96  190]]
```
Accuracy Score:
0.903
AUC:
0.9487413493242508

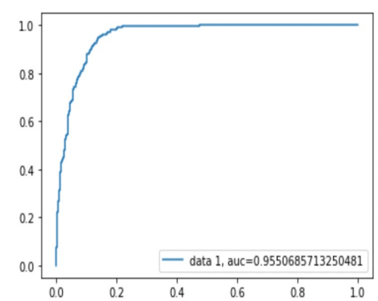


DATA SUBSET USING EXTRA TREES CLASSIFIER

Classification report:

	precision	recall	f1-score	support
0	0.94	0.95	0.95	1544
1	0.73	0.66	0.69	286
accuracy			0.91	1830
macro avg	0.83	0.81	0.82	1830
weighted avg	0.91	0.91	0.91	1830

Confusion matrix:
```
 [[1474   70]
 [  97  189]]
```
Accuracy Score:
0.909
AUC:
0.9550685713250481



Decision Trees Classifier Results

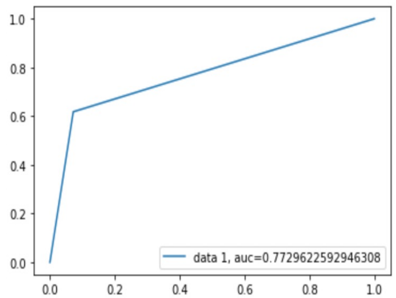
BENCHMARK RESULTS

```
Classification report:
              precision    recall  f1-score   support

           0       0.92      0.93      0.93      1697
           1       0.64      0.62      0.63       348

    accuracy                           0.88      2045
   macro avg       0.78      0.77      0.78      2045
weighted avg       0.87      0.88      0.87      2045

Confusion matrix:
 [[1575  122]
 [ 133  215]]
Accuracy Score:
0.875
AUC:
0.7729622592946308
```



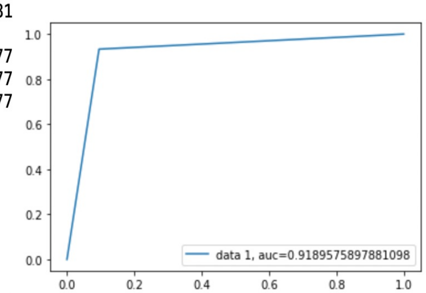
RESAMPLED DATA

```
Classification report:
              precision    recall  f1-score   support

           0       0.94      0.90      0.92      1596
           1       0.90      0.93      0.92      1481

    accuracy                           0.92      3077
   macro avg       0.92      0.92      0.92      3077
weighted avg       0.92      0.92      0.92      3077

Confusion matrix:
 [[1444  152]
 [  99 1382]]
Accuracy Score:
0.918
AUC:
0.9189575897881098
```



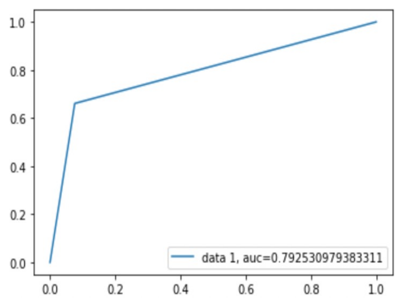
CLEANED PROCESSED DATA

```
Classification report:
              precision    recall  f1-score   support

           0       0.94      0.92      0.93      1544
           1       0.62      0.66      0.64       286

    accuracy                           0.88      1830
   macro avg       0.78      0.79      0.78      1830
weighted avg       0.89      0.88      0.88      1830

Confusion matrix:
 [[1427  117]
 [  97  189]]
Accuracy Score:
0.883
AUC:
0.792530979383311
```



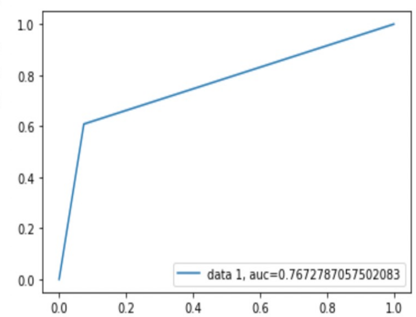
DATA SUBSET USING EXTRA TREES CLASSIFIER

```
Classification report:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93      1544
           1       0.60      0.61      0.61       286

    accuracy                           0.88      1830
   macro avg       0.77      0.77      0.77      1830
weighted avg       0.88      0.88      0.88      1830

Confusion matrix:
 [[1430  114]
 [ 112  174]]
Accuracy Score:
0.877
AUC:
0.7672787057502083
```



Naïve Bayes Classifier Results

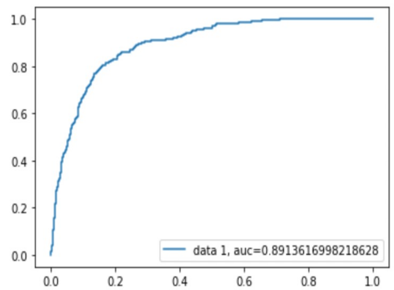
BENCHMARK RESULTS

```
Classification report:
              precision    recall  f1-score   support

           0       0.95      0.83      0.89      1697
           1       0.50      0.80      0.62       348

    accuracy                           0.83      2045
   macro avg       0.73      0.82      0.75      2045
weighted avg       0.88      0.83      0.84      2045

Confusion matrix:
 [[1416  281]
 [  68  280]]
Accuracy Score:
0.829
AUC:
0.8913616998218628
```



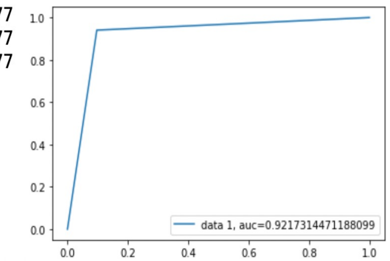
RESAMPLED DATA

```
Classification report:
              precision    recall  f1-score   support

           0       0.94      0.90      0.92      1596
           1       0.90      0.94      0.92      1481

    accuracy                           0.92      3077
   macro avg       0.92      0.92      0.92      3077
weighted avg       0.92      0.92      0.92      3077

Confusion matrix:
 [[1441  155]
 [  88 1393]]
Accuracy Score:
0.921
AUC:
0.9217314471188099
```



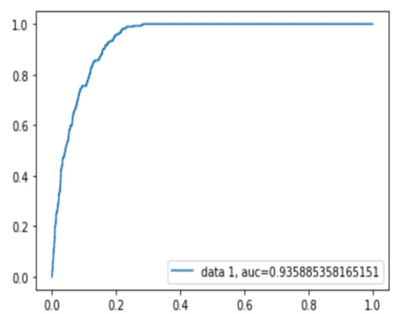
CLEANED PROCESSED DATA

```
Classification report:
              precision    recall  f1-score   support

           0       0.99      0.79      0.88      1544
           1       0.46      0.96      0.62       286

    accuracy                           0.82      1830
   macro avg       0.73      0.88      0.75      1830
weighted avg       0.91      0.82      0.84      1830

Confusion matrix:
 [[1222  322]
 [  11  275]]
Accuracy Score:
0.818
AUC:
0.935885358165151
```



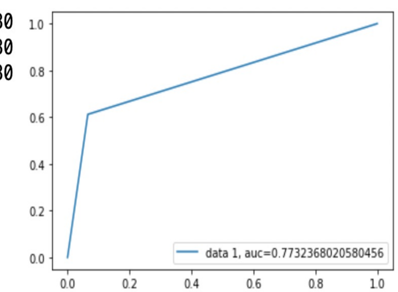
DATA SUBSET USING EXTRA TREES CLASSIFIER

```
Classification report:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93      1544
           1       0.63      0.61      0.62       286

    accuracy                           0.88      1830
   macro avg       0.78      0.77      0.78      1830
weighted avg       0.88      0.88      0.88      1830

Confusion matrix:
 [[1443  101]
 [ 111  175]]
Accuracy Score:
0.884
AUC:
0.7732368020580456
```



K Nearest Neighbors Results (K-3 & K=7)

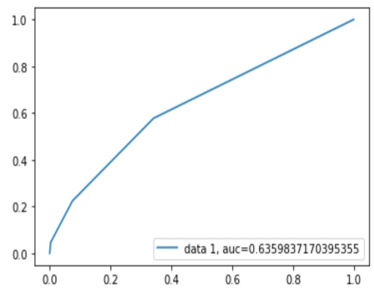
BENCHMARK RESULTS

```
Classification report:
              precision    recall  f1-score   support

           0       0.85      0.93      0.89      1697
           1       0.38      0.22      0.28       348

    accuracy                           0.81      2045
   macro avg       0.62      0.57      0.58      2045
weighted avg       0.77      0.81      0.78      2045

Confusion matrix:
 [[1570  127]
 [ 270   78]]
Accuracy Score:
0.806
AUC:
0.6359837170395355
```



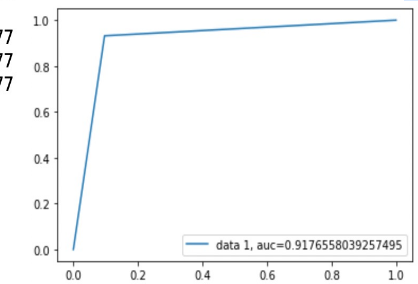
RESAMPLED DATA

```
Classification report:
              precision    recall  f1-score   support

           0       0.93      0.90      0.92      1596
           1       0.90      0.93      0.92      1481

    accuracy                           0.92      3077
   macro avg       0.92      0.92      0.92      3077
weighted avg       0.92      0.92      0.92      3077

Confusion matrix:
 [[1442  154]
 [ 101 1380]]
Accuracy Score:
0.917
AUC:
0.9176558039257495
```



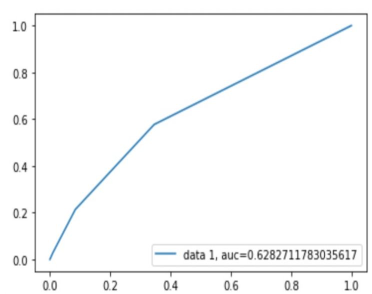
CLEANED PROCESSED DATA

```
Classification report:
              precision    recall  f1-score   support

           0       0.86      0.92      0.89      1544
           1       0.32      0.21      0.26       286

    accuracy                           0.81      1830
   macro avg       0.59      0.56      0.57      1830
weighted avg       0.78      0.81      0.79      1830

Confusion matrix:
 [[1414  130]
 [ 225   61]]
Accuracy Score:
0.806
AUC:
0.6282711783035617
```



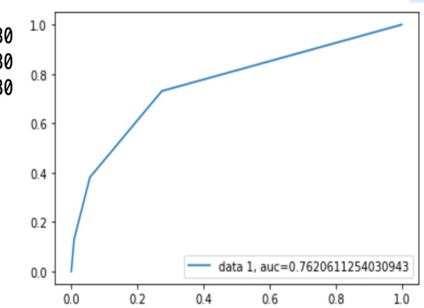
DATA SUBSET USING EXTRA TREES CLASSIFIER

```
Classification report:
              precision    recall  f1-score   support

           0       0.89      0.94      0.92      1544
           1       0.56      0.38      0.45       286

    accuracy                           0.86      1830
   macro avg       0.72      0.66      0.68      1830
weighted avg       0.84      0.86      0.84      1830

Confusion matrix:
 [[1457   87]
 [ 177  109]]
Accuracy Score:
0.856
AUC:
0.7620611254030943
```



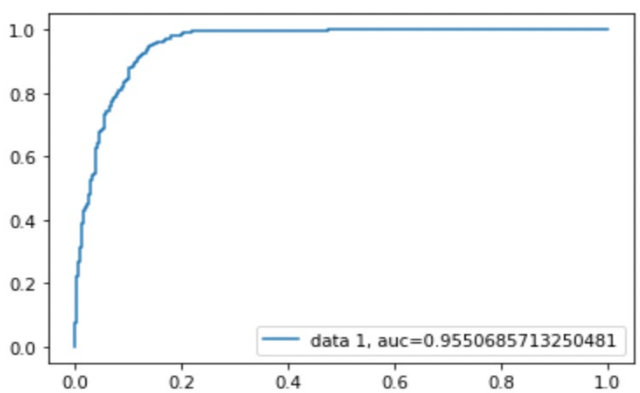
Selected Model & Results

```
Classification report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.95      1544
           1       0.73      0.66      0.69       286

    accuracy                           0.91      1830
   macro avg       0.83      0.81      0.82      1830
weighted avg       0.91      0.91      0.91      1830

Confusion matrix:
 [[1474   70]
 [  97  189]]
Accuracy Score:
0.909
AUC:
0.9550685713250481
ROC:
```



Winning Model : Logistic regression

Winning Dataset : Data Subset Using decision trees Classifier

Top features :

'int_corr' : correlation between participant's and partner's ratings of interests

'age_o' : age of partner

'pf_o_att' , 'pf_o_int', 'pf_o_fun', 'pf_o_amb', 'pf_o_sha' : Partner's stated preference for attraction, intelligence, fun, ambition, shared interest

'attr_o', 'sinc_o', 'intel_o', 'fun_o', 'amb_o', 'shar_o' : Rating by partner the night of the event, for attraction, intelligence, fun, ambition, shared interest

'like_o' : partners rating of like

'dec' : decision of participant

'attr' : attractiveness of partner

'intel', 'fun', 'amb', 'shar', 'like' : Participants rating on intelligence, fun, ambition, shared interest and like

'prob' : How probable do you think it is that this person will say 'yes' for you

Q & A