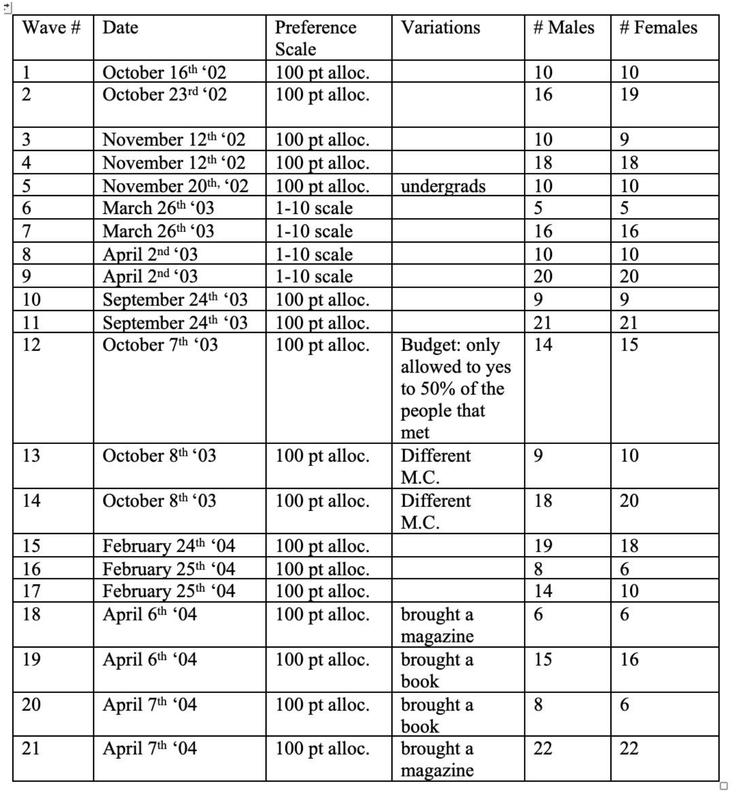
5243 ADS – Project 5

Speed Dating Analysis :

Predicting the likelihood of having a follow-up date

Yerin Cho, Rhoan Lee, Licheng Wu, Christine Lu

Dataset Summary



Wave #	Date	Preference Scale	Variations	# Males	# Females
1	October 16th ‘02	100 pt alloc.		10	10
2	October 23rd ‘02	100 pt alloc.		16	19
3	November 12th ‘02	100 pt alloc.		10	9
4	November 12th ‘02	100 pt alloc.		18	18
5	November 20th, ‘02	100 pt alloc.	undergrads	10	10
6	March 26th ‘03	1-10 scale		5	5
7	March 26th ‘03	1-10 scale		16	16
8	April 2nd ‘03	1-10 scale		10	10
9	April 2nd ‘03	1-10 scale		20	20
10	September 24th ‘03	100 pt alloc.		9	9
11	September 24th ‘03	100 pt alloc.		21	21
12	October 7th ‘03	100 pt alloc.	Budget: only allowed to yes to 50% of the people that met	14	15
13	October 8th ‘03	100 pt alloc.	Different M.C.	9	10
14	October 8th ‘03	100 pt alloc.	Different M.C.	18	20
15	February 24th ‘04	100 pt alloc.		19	18
16	February 25th ‘04	100 pt alloc.		8	6
17	February 25th ‘04	100 pt alloc.		14	10
18	April 6th ‘04	100 pt alloc.	brought a magazine	6	6
19	April 6th ‘04	100 pt alloc.	brought a book	15	16
20	April 7th ‘04	100 pt alloc.	brought a book	8	6
21	April 7th ‘04	100 pt alloc.	brought a magazine	22	22

Feature Summary

Category	Details
Participant Basics	- Participant’s unique identifier (iid) - Demographic information: gender, age, race, educational background, and career intentions
Event Details	- The date and the method used for preference evaluation (100-point distribution or a 1-10 scale)
Mutual Ratings	- Participants’ ratings based on attractiveness, sincerity, and intelligence of their dates - How they believe they were rated by their counterparts.
Decisions and Preferences	- Whether they want to meet each person again after each date and assess their own and their counterparts' preferences.
Post-Event Evaluation	- Participants evaluate their satisfaction with the event, the number of dates, and the characteristics of the dates during and after the event.

Key Features of Interest

Key Features

‘match’ 1=yes, 0=no

‘date_3’:

Have you been on a date with any of your matches?

Yes=1

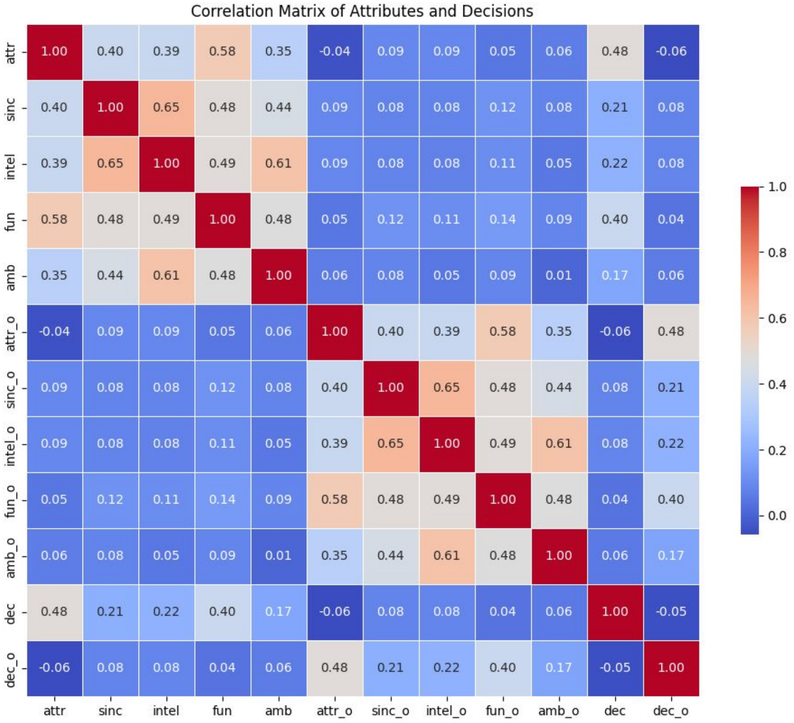
No=2

If you have been on at least one date, please answer the following:

‘numdat_3’:

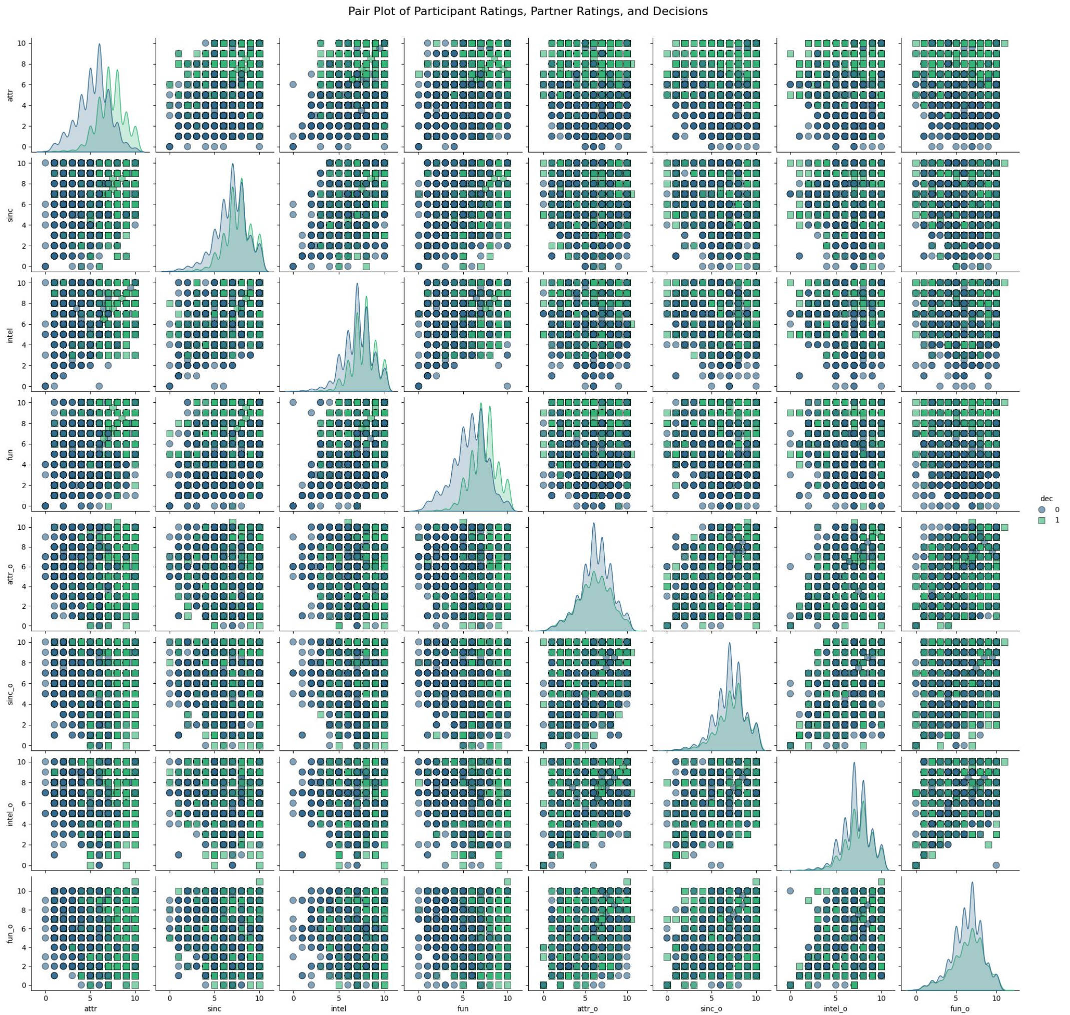
(a) How many of your matches have you been on a date with so far?

EDA



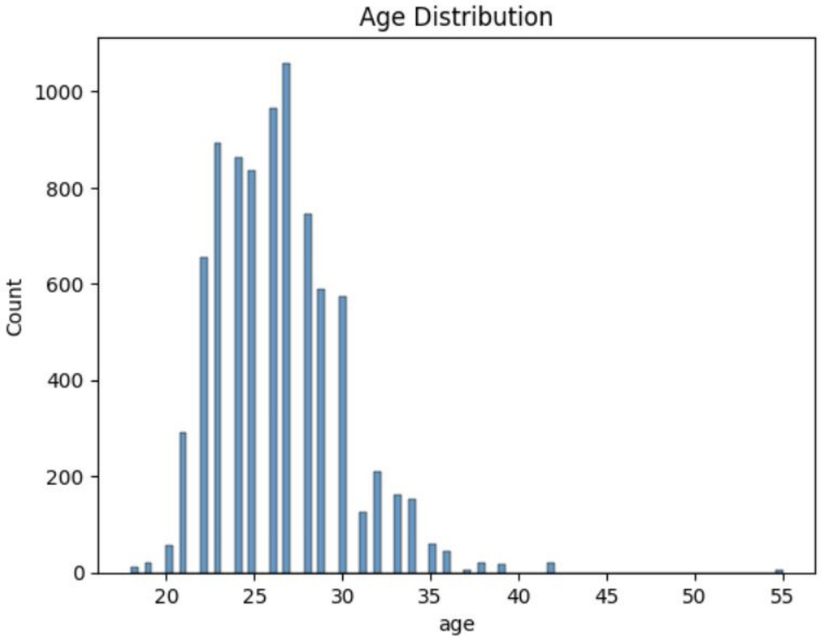
Correlation Matrix of Attributes and Decisions

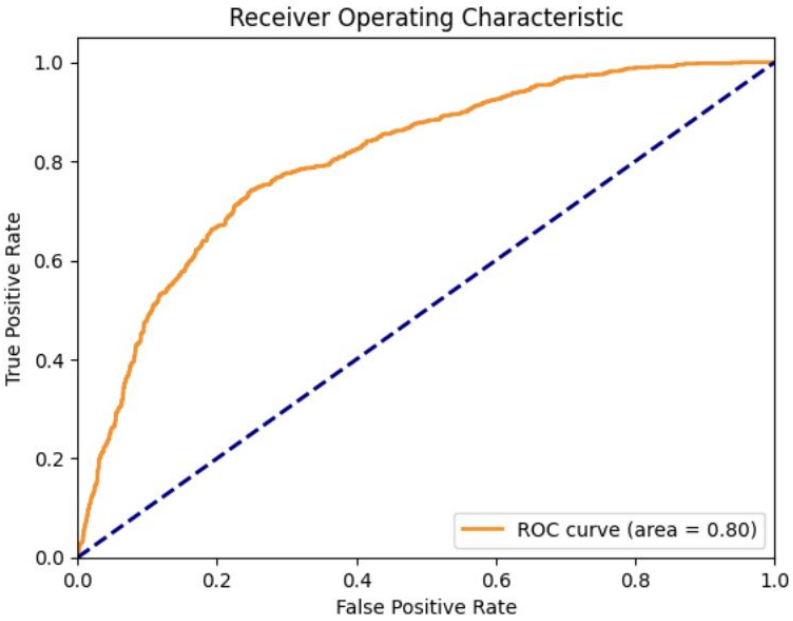
	attr	sinc	intel	fun	amb	attr_o	sinc_o	intel_o	fun_o	amb_o	dec	dec_o
attr	1.00	0.40	0.39	0.58	0.35	-0.04	0.09	0.09	0.05	0.06	0.48	-0.06
sinc	0.40	1.00	0.65	0.48	0.44	0.09	0.08	0.08	0.12	0.08	0.21	0.08
intel	0.39	0.65	1.00	0.49	0.61	0.09	0.08	0.08	0.11	0.05	0.22	0.08
fun	0.58	0.48	0.49	1.00	0.48	0.05	0.12	0.11	0.14	0.09	0.40	0.04
amb	0.35	0.44	0.61	0.48	1.00	0.06	0.08	0.05	0.09	0.01	0.17	0.06
attr_o	-0.04	0.09	0.09	0.05	0.06	1.00	0.40	0.39	0.58	0.35	-0.06	0.48
sinc_o	0.09	0.08	0.08	0.12	0.08	0.40	1.00	0.65	0.48	0.44	0.08	0.21
intel_o	0.09	0.08	0.08	0.11	0.05	0.39	0.65	1.00	0.49	0.61	0.08	0.22
fun_o	0.05	0.12	0.11	0.14	0.09	0.58	0.48	0.49	1.00	0.48	0.04	0.40
amb_o	0.06	0.08	0.05	0.09	0.01	0.35	0.44	0.61	0.48	1.00	0.06	0.17
dec	0.48	0.21	0.22	0.40	0.17	-0.06	0.08	0.08	0.04	0.06	1.00	-0.05
dec_o	-0.06	0.08	0.08	0.04	0.06	0.48	0.21	0.22	0.40	0.17	-0.05	1.00

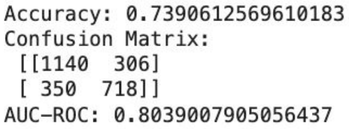


Pair Plot of Participant Ratings, Partner Ratings, and Decisions

EDA

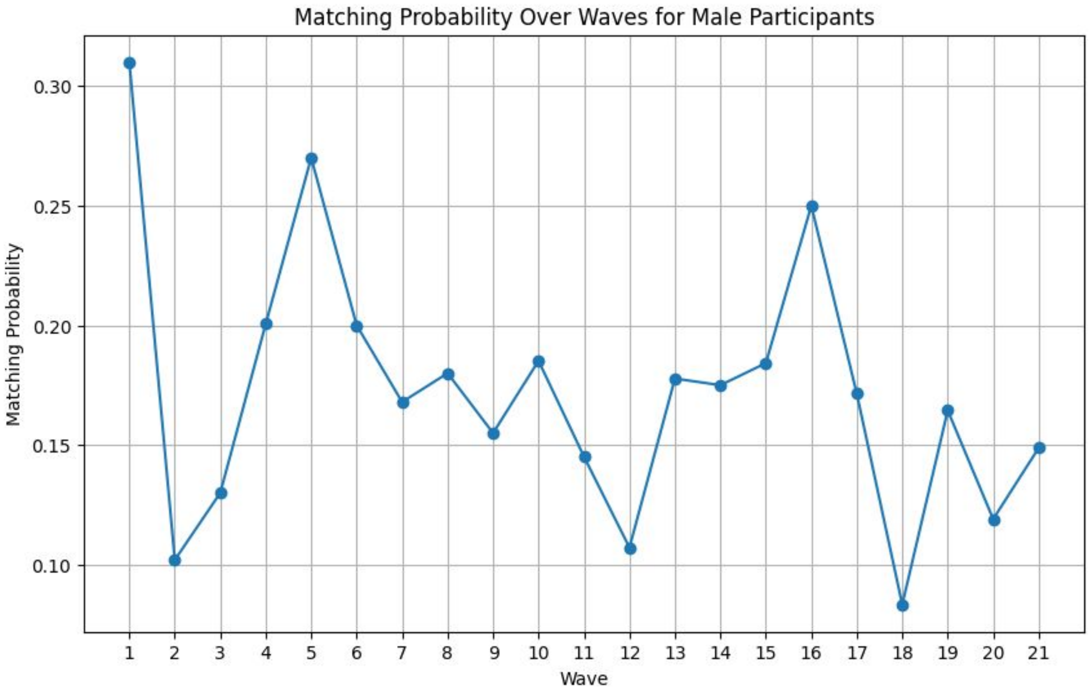






```
Accuracy: 0.7390612569610183
Confusion Matrix:
 [[1140  306]
 [ 350  718]]
AUC-ROC: 0.8039007905056437
```

Temporal Analysis



Methodologies

Logistic: Baseline Model

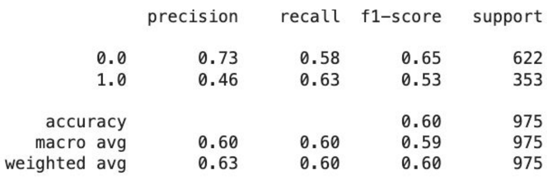
Random Forest: Random Forest is chosen for its proficiency in handling complex datasets with many features, capable of capturing non-linear relationships. The ensemble approach, combining multiple decision trees, reduces the risk of overfitting and provides a more generalized model. Random Forest can effectively handle imbalanced datasets, which is common in long-term outcome predictions where one result (like no follow-up date) may dominate.

→ Targeted analysis for ‘date_3’ (whether there was a follow up date)

XGBoost: XGBoost is chosen to build a model using the provided features and evaluates its accuracy on the test data. XGBoost has many parameters that can be tuned for better performance, but this code uses default parameters for simplicity.

Predicting the likelihood of having a follow-up date − **Results**

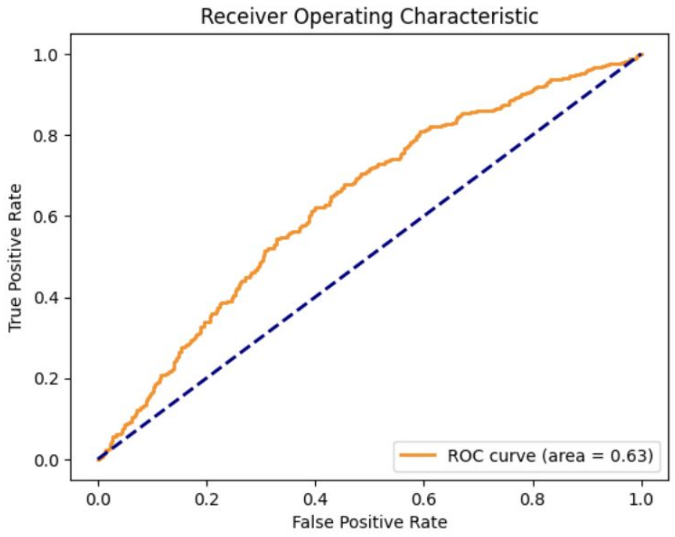
Logistic



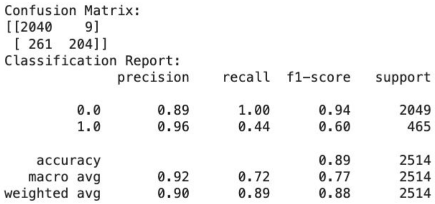
```
              precision    recall  f1-score   support

         0.0       0.73      0.58      0.65       622
         1.0       0.46      0.63      0.53       353

    accuracy                           0.60       975
   macro avg       0.60      0.60      0.59       975
weighted avg       0.63      0.60      0.60       975
```



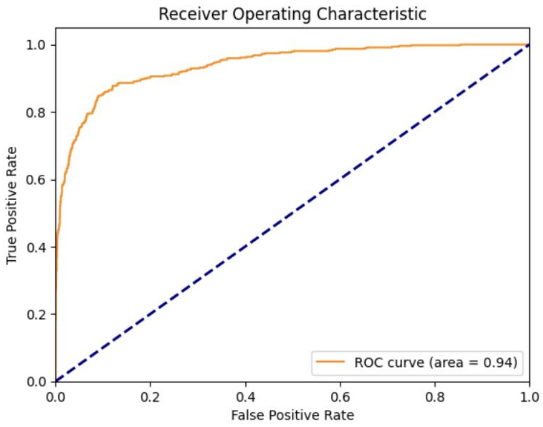
Random Forest



```
Confusion Matrix:
[[2040    9]
 [ 261  204]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.89      1.00      0.94      2049
         1.0       0.96      0.44      0.60       465

    accuracy                           0.89      2514
   macro avg       0.92      0.72      0.77      2514
weighted avg       0.90      0.89      0.88      2514
```



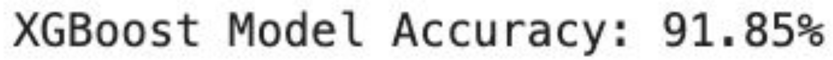
The model achieves an overall accuracy of 89.26%, which is excellent but may be somewhat inflated due to class imbalance.

Targeted analysis on where there was a follow-up date

Gradient Boosting

Based on the logistic and Random Forest model, we built an XGBoost model to predict whether there was a follow up date.

For this model, we defined ‘date_3’ as the objective feature.

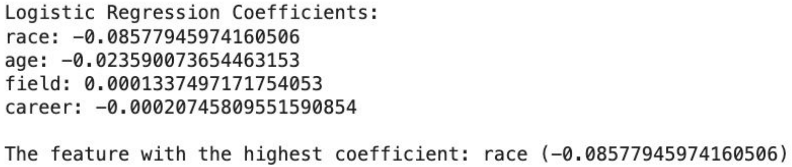


```
XGBoost Model Accuracy: 91.85%
```

Which attribute has the highest influence on male's matching?

Determine which among race, age, field of study, and career has the highest influence on male's matching

- Use **logistic regression coefficients** as a measure of importance



```
Logistic Regression Coefficients:
race: -0.08577945974160506
age: -0.023590073654463153
field: 0.0001337497171754053
career: -0.00020745809551590854

The feature with the highest coefficient: race (-0.08577945974160506)
```

- How?

  First, fit a logistic regression model to predict match outcomes for males using race, age, field of study, and career as independent variables.

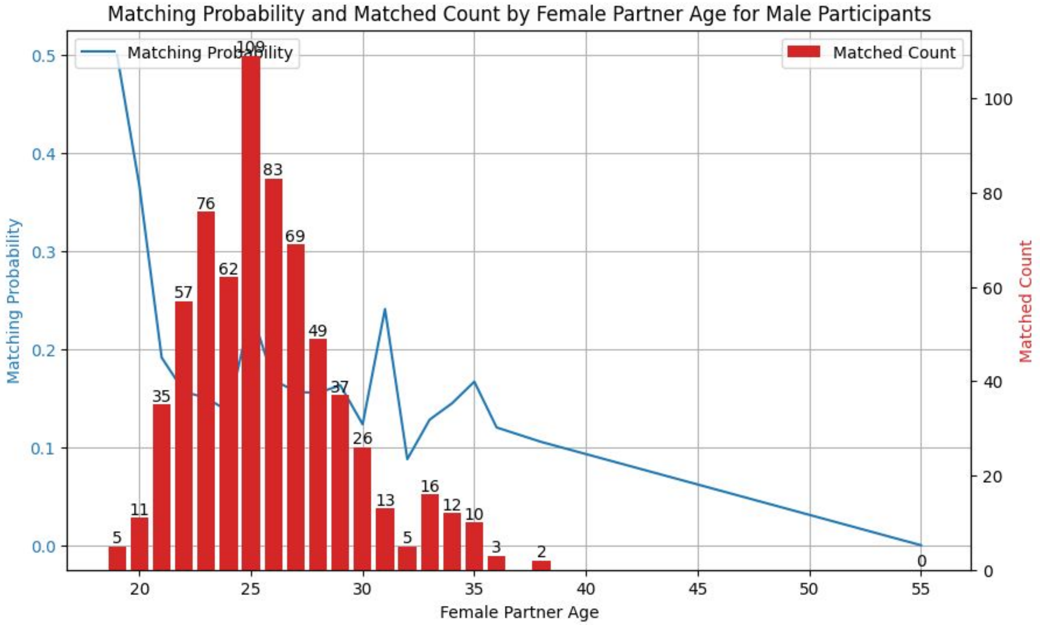
  Then extract the coefficients of the logistic regression model.

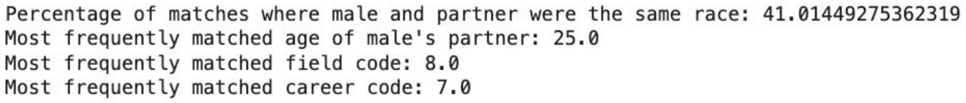
  Lastly analyze the magnitudes of the coefficients to determine which variable has the highest influence on match outcomes for males.

Results & Conclusion

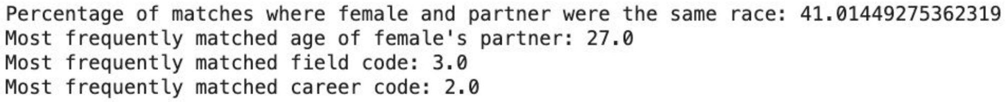
- field_cd: field coded
1= Law
2= Math
3= Social Science, Psychologist
4= Medical Science, Pharmaceuticals
5= Engineering
6= English/Creative Writing/ Journalism
7= History/Religion/Philosophy
8= Business/Econ/Finance
9= Education, Academia
10= Biological Sciences/Chemistry/Physics
11= Social Work
12= Undergrad/undecided
13=Political Science/International Affairs
14=Film
15=Fine Arts/Arts Administration
16=Languages
17=Architecture
18=Other

- career_c: career coded
1= Lawyer
2= Academic/Research
3= Psychologist
4= Doctor/Medicine
5=Engineer
6= Creative Arts/Entertainment
7= Banking/Consulting/Finance/Marketing/Business
8= Real Estate
9= International/Humanitarian Affairs
10= Undecided
11=Social Work
12=Speech Pathology
13=Politics
14=Pro sports/Athletics
15=Other
16=Journalism
17=Architecture





```
Percentage of matches where male and partner were the same race: 41.01449275362319
Most frequently matched age of male's partner: 25.0
Most frequently matched field code: 8.0
Most frequently matched career code: 7.0
```



```
Percentage of matches where female and partner were the same race: 41.01449275362319
Most frequently matched age of female's partner: 27.0
Most frequently matched field code: 3.0
Most frequently matched career code: 2.0
```