

**Will I Want You to Be  
My Date?**



# Problem Statement

## Background:

- Experimental study conducted at Columbia Business School
- Data from speed dating events from 2002-2004
- Participants interact with potential romantic partners of the opposite sex for four minutes and then answer questions about the it and decide if they will want to go on a date or not

## Question:

- Predict if a person in the study would decide to go on a date with the newly met individual.

## Response variable:

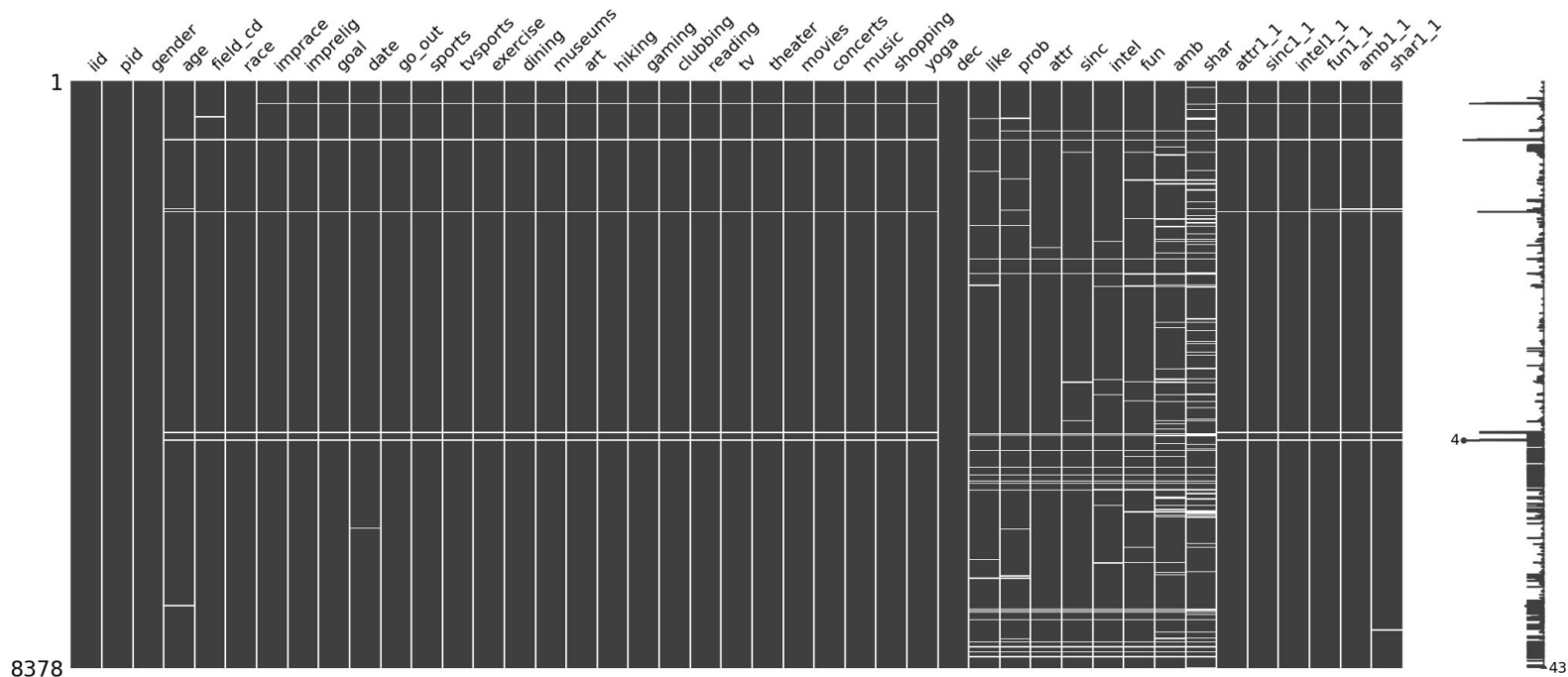
- “Dec”: decision whether the participants would want to go on a date with the potential partner or not.

# Data Overview

- 551 participants: 50.3 % male and 49.7% female
- **8 368 two people interactions**
- From original 195 features, 43 features were chosen to be analyzed
- Created additional variables of difference between participant and potential partner
- Four big categories of features:
  - Demographic information
  - Interest in leisure activities
  - Perception of the other person
  - Importance that each of the participants assigns to each of the features

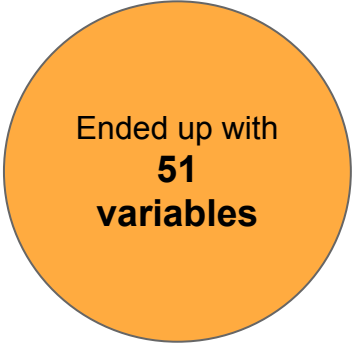
# Data cleaning & transformation

- Out of the 43 features we were interested in, 40 had missing observations



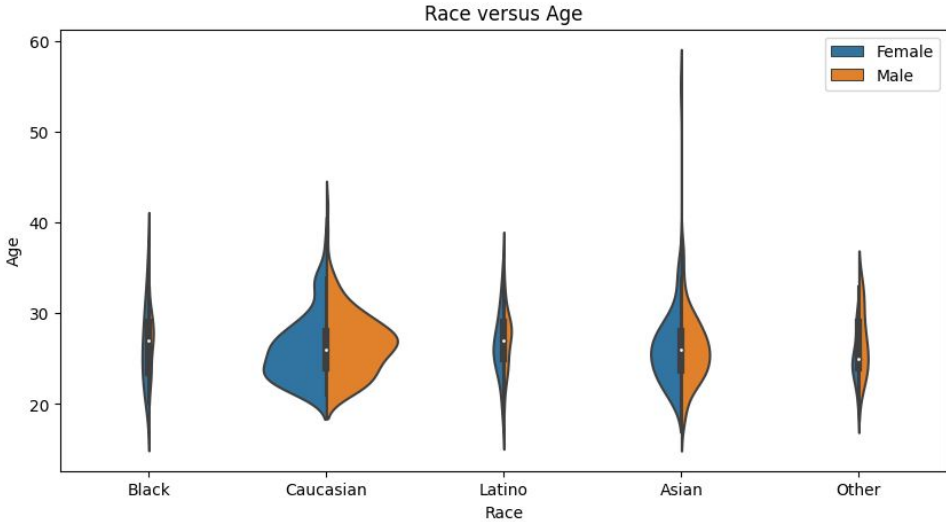
# Data cleaning & transformation

- Missingness imputation
  - Created unique dataframe of participants
  - For features that don't depend on partner:
    - For numerical: kNN (k=5) imputation
    - For categorical: Mode imputation
  - Merge original dataset with unique dataset
  - Impute features that depend on partner on entire dataset
- Data cleaning
  - Rescaled inconsistencies
  - Delete observations with missing partner's ID
- Feature engineering
  - Consolidate high cardinality categorical features.
  - Create additional variables: difference for numerical and comparison for categorical.
  - Apply Variance Inflation Factor (VIF) to address collinearity
  - Remove non-relevant features like the IDs
  - Apply One-Hot-Encoding for categorical variables



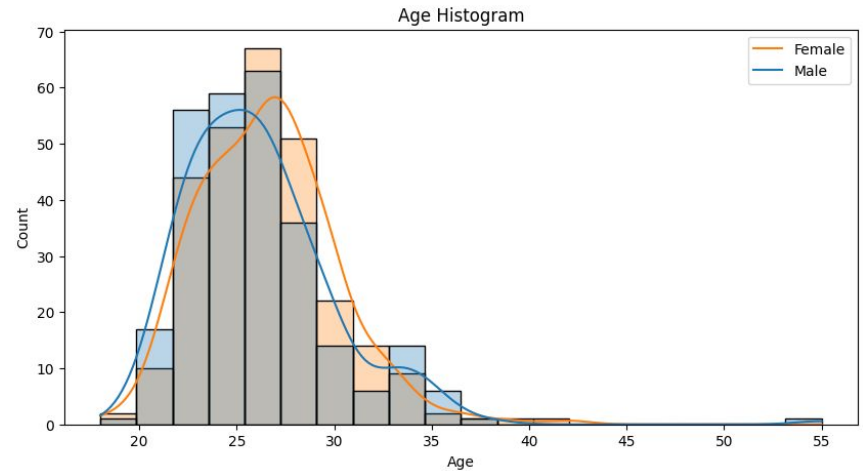
Ended up with  
**51**  
**variables**

# Exploratory Data Analysis



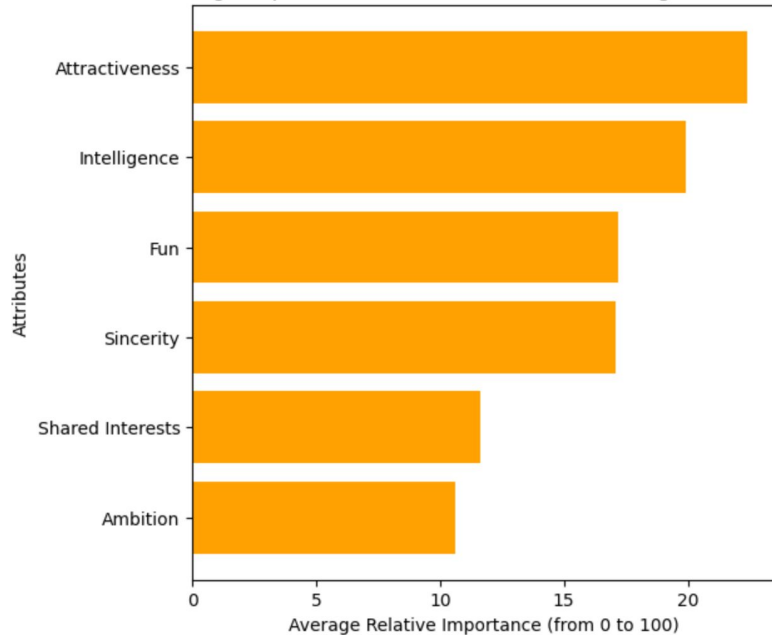
Most participants were Caucasian, followed by Asians.

Most participants had an age between 20 and 30

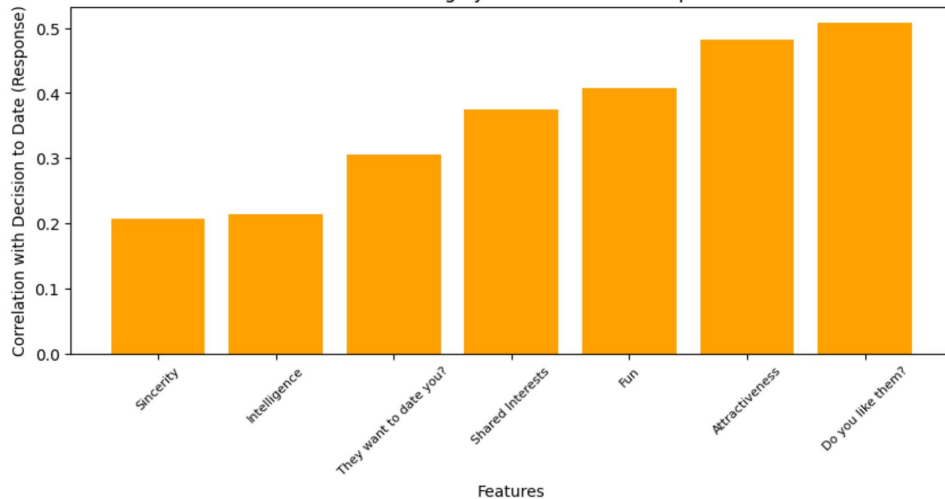


# Exploratory Data Analysis

Average Importance of Desired Attributes In a Significant Other



Features Most Highly Correlated With Response



# Models for classification

Baseline  
Model

Logistic  
Regression

KNN

Decision  
Tree

Random  
Forest

XGBoost

Advanced Section model:

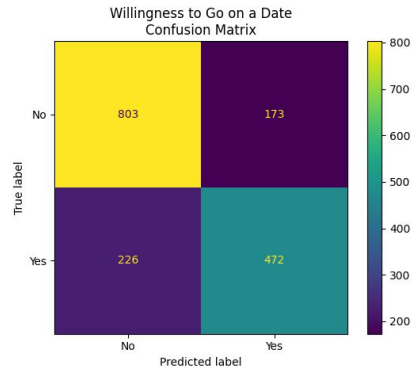
Stacking



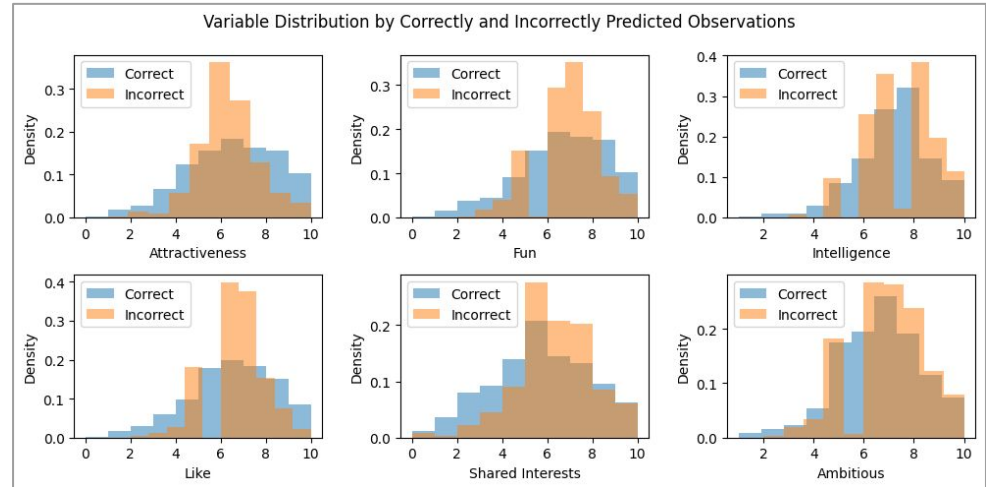
# Baseline Model

Performed logistic regression model without cross-validation or penalty.

- **Train accuracy: 79.24%**
- **Test accuracy: 76.7%**



Baseline Model Most Important Features		
	Feature	Odds Change
1.	Overall, how much do you like this person?	2.72
2.	Rate attractiveness of potential date	2.36
3.	Are you male?	1.66
4.	How probable do you think it is that this person will say “yes” for you?	1.47
5.	Rate “funness” of potential partner	1.32



# Lasso Logistic Regression

- 40 important, 10 non-important features.

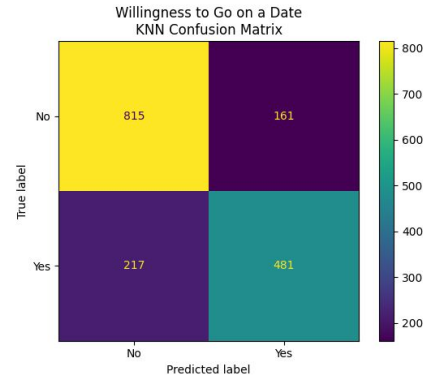
- **Train accuracy:** 79.18%
- **Test accuracy:** 76.17%

## KNN

- 30 neighbors.

- **Train accuracy:** 100%
- **Test accuracy:** 77.42%

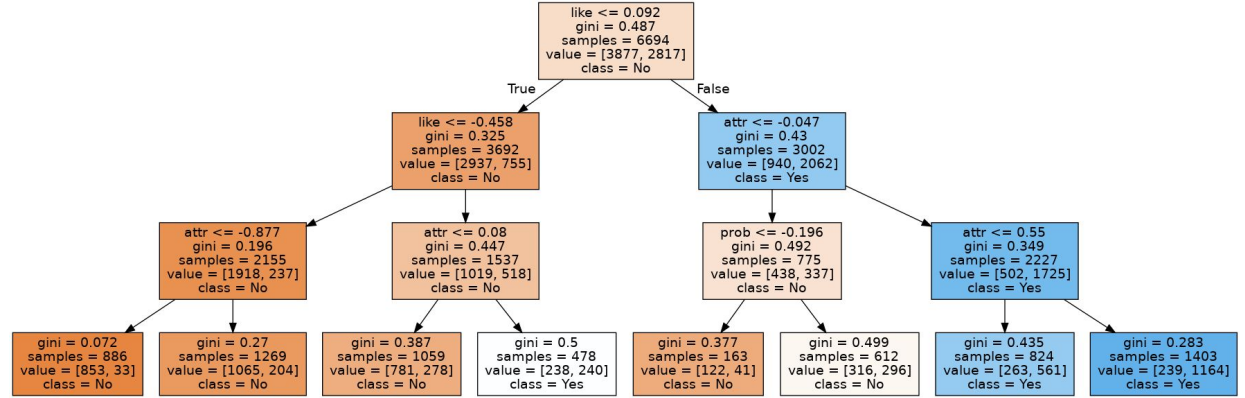
Lasso Model Non-Important Features
Field of study (OHE variables)
Do you and potential partner go out with the same frequency? (Not necessarily on dates)
Difference in importance given to dating someone of the same race
Rate intelligence of potential partner
Difference in interest in theater and yoga



# Decision Tree

- Max depth: 8
- Splitter: best

- **Train accuracy: 85.36%**
- **Test accuracy: 77%**



Decision Tree Most Important Features

1.	Overall, how much do you like this person?	0.50
2.	Rate attractiveness of potential date	0.12
3.	How probable do you think it is that this person will say "yes" for you?	0.03
4.	How important it is for you that a potential date has shared interests / hobbies?	0.03
5.	How important it is for you that a potential date is fun?	0.03

# Increasing the Model Complexity of a Single Decision Tree

**XGBoost** *Sequentially fitting different weak decision trees and learning from mistakes.*

The Loss function is regularized using l1 and l2:  
includes a penalty for model complexity

$$\mathcal{L}^t = \sum_{i=1}^n l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t)$$

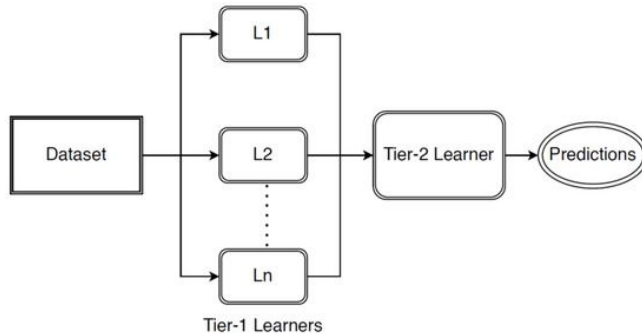
- **Train accuracy:** 100%
- **Test accuracy:** 82.02%

**Random Forest** *Combining the predictions of multiple de-correlated trees*

- **Best parameters:**
  - Max depth: 20
  - Estimators: 600
- **Train accuracy:** 100%
- **Test accuracy:** 79.57%

# Stacking

Stack the methods run before: Logistic regression, KNN, Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier as base learners



- **Train accuracy:** 100%
- **Test accuracy:** 82.08%

# Results

# Results

Decision Trees chosen as best models because:

- Easier to interpret
- Being someone that other people enjoy being around (being a likable person) can boost one's chances of success in dating

Model	Train Accuracy	Test Accuracy
Baseline: Logistic Regression	0.7918	0.7616
Logistic Regression	0.7924	0.7670
KNN	1.0	0.7742
Decision Tree	0.8536	0.7700
Random Forest	1.0	0.7957
XGBoost	1.0	0.8202
Stacking	1.0	0.8208