### **Project Problem and Hypothesis**

* *What's the project about? What problem are you solving?*

Two goals of this project:

1. To predict the probability of matching in speed dating based on shared interest, age difference, same or different race, difference in attractiveness, intelligence, ambition between the pair.
2. Breaking up by gender, to predict the probability of either gender’s decision to go on a date with the other person based on ratings on the other person’s attractiveness, sincerity, intelligence, fun, ambition and shared interested.

* *Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?*

It could be binary value, match/non match, decision to go on a date/no date. I will also be interested in using logistic regression to predict probability, which will be a continuous number between 0 and 1.

* *What kind of impact do you think it could have?*

The research digs into human behavior on dating based on gender, and race. It will allow us to have a general understanding of modern day dating behavior in the U.S.

* *What do you think will have the most impact in predicting the value you are interested in solving for?*

For the first goal: shared interest and attractiveness probably have larger impact in predicting probability of match.

For the second goal: attractiveness and intelligence probably have larger impact on participant’s decision to go on a date.

### **Datasets**

* *Description of data set available, at the field level (see table)*

Dataset is downloaded from Kaggle.com. Participants are students in graduate and professional schools at Columbia University. Participants were recruited through a combination of mass e-mail and fliers posted throughout the campus and handed out by research assistants.

There are over 190 fields that were collected in the original dataset. Below are the field that I selected for my analysis:

|  |  |
| --- | --- |
| iid | Unique subject number in a group |
| id | Subject number within group |
| gender | 0: female; 1: male |
| match | 1: yes; 0: no |
| int\_corr: | Correlation between participant’s and the partner’s ratings of interests |
| samerace | 1: yes; 0: no |
| age\_o: | Age of partner |
| attr\_o: | Attractiveness rated by partner on the night |
| int\_o | Intelligence rated by partner on the night |
| amb\_o | Ambitiousness rated by partner on the night |
| age: | Participant’s age |
| race: | 1: Black/African american  2: European/Caucasian American  3: Latino/Hispanic American  4: Asian/Pacific Islander/Asian-American  5: Native American  6: Other |
| attr3\_1 | Participant’s opinions on their own attractiveness. On scale 1-10, 1 is awful, 10 is great. |
| int3\_1 | Participant’s opinions on their own intelligence. On scale 1-10, 1 is awful, 10 is great. |
| amb3\_1 | Participant’s opinions on their own ambitiousness. On scale 1-10, 1 is awful, 10 is great. |
| dec | Participant’s decision if they would go on a date with the other person: 1: yes, 0: no |
| attr | Participant’s ratings of the other person’s attractiveness |
| intel | Participant’s ratings of the other person’s intelligence |
| amb | Participant’s ratings of the other person’s ambitiousness |
| Shared interests/hobbies | Participant’s ratings of the shared common interests with the other person |
| prob | How probably participants think the other person will say ‘yes’ |

### **Domain knowledge**

* What experience do you already have around this area?

Personal experience and common sense. Some knonwlege of decision science and human behavior.

* Does it relate or help inform the project in any way?

Yes, it helped me to select features that are relevant to predict target.

* What other research efforts exist?

Upon research on the Internet, there’s no much existing research on this topic. There’re some articles on eHarmony and okTrend from okcupid.

* + *Use a quick Google search to see what approaches others have made, or talk with your colleagues if it is work related about previous attempts at similar problems.*
  + *This could even just be something like "the marketing team put together a forecast in excel that doesn't do well."*
  + *Include a benchmark, how other models have performed, even if you are unsure what the metric means.*

### **Project Concerns**

* *What questions do you have about your project? What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help).*

I’m not sure yet. I haven’t gotten chance to dig into the data.

* *What are the assumptions and caveats to the problem?*
  + *What data do you not have access to but wish you had?*

It’d be great to have sample data from general public, meaning a mix of education background, different sexual orientation (LGBT), and geographic areas.

* + *What is already implied about the observations in your data set? For example, if your primary data set is twitter data, it may not be representative of the whole sample (say, predicting who would win an election)*

Data was collected from Columbia University students so this represents a specific subset of population: educated, middle-class American. Although not specifically indicated, the speed datings were set up for heterosexual couples only.

* *What are the risks to the project?*
  + *What's the cost of your model being wrong? (What's the benefit of your model being right?)*

Cost of model being wrong: incorrectly predicts matching probability or a person’s decision on going on a date.

Benefit of model being right: correctly predicts matching probability given two people race, age difference, and ratings difference on 4 attributes.

* + *Is any of the data incorrect? Could it be incorrect?*

Data was collected from self-reported questionnaire. There could be mistakes in filling out the questionnaire or participants could intentionally fill in false information. For example, attractiveness score on themselves.

### **Outcomes**

* *What do you expect the output to look like?*

I expect to fit the data into a logistic regression model, and discover significant relationships between certain features and target.

* *What does your target audience expect the output to look like?*

My target audience probably expects a story of modern day dating or a magic 8-ball to predict a matching score based on attributes of two people.

* *What gain do you expect from your most important feature on its own?*

I expect to see different decision-making behaviors between women/man, between different races, between high self-esteem vs. low self-esteem.

* *How complicated does your model have to be?*

It doesn’t have to be complicated. A few features (based on common sense) are included and to be tested out. Features I’m considering are: gender, race, age, shared interest.

* *How successful does your project have to be in order to be considered a "success"?*

I consider the project to be successful if I discover some of the features I chose are significant predictors to target. Or if there’s any significant correlation found between features.

* *What will you do if the project is a bust (this happens! but it shouldn't here)?*

I will try to modify the features I chose. I do believe there’s fundamental logic in people’s decision in choosing potential dates. Therefore, there should be some kind of relationship to be discovered in this dataset.