Project Problem and Hypothesis

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

* The project is about trying to understand what factors are important in selecting a partner in a speed dating situation.

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

* Binary value: Can we predict if a person is going to choose to see someone again based on known information about both individuals prior to meeting?

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

* Improve dating matching algorithms, to introduce two people who are likely to be interested in each other.

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

* Perception of self across 6 attributes
* Attributes of interest in the opposite sex across 6 attributes
* Self-perception of the partner across 6 attributes
* Race
* Religion
* Attitude about the speed dating event

Datasets

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

* Columbia speed dating data set.
* <<Speed Dating Data Key (1).doc>>

Domain knowledge

•What experience do you already have around this area?

General ideas of social interaction and social sciences

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

•What other research efforts exist? ◦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.

 Existing research:

* + “Racial Preferences in Dating: Evidence from a Speed Dating Experiment”
    - Used multiple models: linear regression probability and OLS models. Rsquared of up to ~0.50.
  + “Gender Differences in Mate Selection”
    - Linear probability model. R squared of up to ~0.50.

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)

* + There are many relevant variables in this dataset. The feature engineering process will be complex, especially because variables will have to be considered for both individuals in the date. Should I consider the attributes for each individual separately and on an absolute basis, or relative to each other? Or should I consider it relative to what the dater has stated their preferences are?

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

* + I wish I had information on whether or not the relationship worked out. I also wished that each date was longer than 4 minutes. The 4-minute limit may not be enough time to assess none superficial attributes that may not be obvious within a 4-minute meeting. This skews results to more superficial attributes such as looks.

◦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)

* + Assuming that the sample population are all Columbia students; may not be reflective of the entire population

•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?)

◦Is any of the data incorrect? Could it be incorrect?

* + The risk of the model being wrong is low. There has not be many ‘scientific’ approaches to research in this field, so any research / development is incremental

Outcomes

•What do you expect the output to look like?

* + I expect to create a model based on known variables prior to the date, and predict if the date will be successful or not.

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

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

•How complicated does your model have to be?

* + I don’t think the model itself will be too complex once the set of features have been defined. I may have to create multiple models for each gender and / or race, depending on how different their behaviors are.

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

* + I would like to be able to correctly predict the outcome of >75% dates.

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

* + Seek other datasets to explore other angles to this question that might not be enabled with this dataset