**Capstone Project**

Jesse Gallegos  
January 12th, 2017

**I. Definition**

**Project Overview**

This project investigates a Speed Dating dataset from [Kaggle](https://www.kaggle.com/annavictoria/speed-dating-experiment). The dataset was compiled by two professors from Columbia University's Business School. Information including demographics and preferences of attraction were collected from participants during weeknights in the years 2002 through 2004. Moreover, the study aimed at reproducing a business model that Speed Dating services use and chose similar environments to hold these events (e.g. popular bars near universities).

During the events, daters had a four minute "first date" with every other participant of the opposite sex. At the end of each four minute "date", participants were asked if he, or she, would date that person again. Participants were also asked to rate his, or her, date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests.

In addition, the dataset includes a questionnaire that gathered data from participants at different points in the process. These fields include: demographics, dating habits, self-perception. Other data include what features of attraction participants think people find valuable in a mate. These preferences of attraction were asked at the start of the event, half way through the event, a day after the event and three weeks after the event.

Finally, the dataset yielded a paper, Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment, by professors Ray Fisman and Sheena Iyengar. The rest of project will explore the mentioned dataset.

**Problem Statement**

One finding to verify with the papers written by the Columbia professors is that men value physical attractiveness of a partner more than women do. Another is that women place a greater weight on intelligence than men do. Finally, this project will do two things:

1. The first task at hand is to get an intuitive sense of the data. The goal is to build visuals that summarize key aspects of the dataset.
2. The other task is to create a model to predict whether a person decides to go on a second with the other person.

A 'match' in the given dataset is defined as a man and a woman having decided, 'dec,' 'yes' to date each other again. In short, 'match' depends on 'dec' of a male and 'dec' of a female. There are instances in the dataset where a person of either gender say 'yes,' but the other party says 'no.' By definition, those are not matches.

**Metrics**

The main challenge is dealing with how my model assigns a 'yes' or 'no' decision. The algorithm could incorrectly label a person saying 'no' to go out with the other person again when in fact the decision was 'yes'. The model could also incorrectly assign a 'yes' label' when the person actually decided 'no'. The two other cases are the model outputting 'yes' when a person wants a second date. Likewise, the model could output 'no' when a respondent does not want a second date. These four possibilities make precision and recall computable. Precision is important to the model because the model needs to correctly predict second dates. Recall is important to the model because it measures how relevance (in a future project this is important for match making). f1score is the ideal performance metric for those reasons.

At first, my evaluation metric was Mean Absolute Error (MAE). The goal was to have MAE measure the error between what I am doing in this project against the results published by Fisman and Iyengar. This evaluation metric would have compared the total sum of decisions against the number of decisions of 'yes' decisions the professors predicted. After going through the benchmark model, I realized that the authors of the article incorrectly used a regression algorithm for classification. However, in spite of that, the two professors did publish basic statistics, standardized their chosen features. This will aid in doing sanity checks in lieu of the model chosen by the authors.

**II. Analysis**

*(approx. 2-4 pages)*

**Data Exploration**

The dataset file, Speed Dating Data, has 8378 data points with 195 different features. There were 551 people involved in the dating event with the population being divided into 274 women and 277 men. For the sake of brevity, the reader is invited to consult the Data Key file included with the dataset. The Key file gives the exact number of participants for each of the 21 'wave' events. Moreover, the Data Key file gives a thorough overview for what the allowed values for each field are.

The first attributes that will be explored at the start of the project include 'iid', 'gender', 'wave', 'attr', 'intel', 'sinc', 'fun', 'amb', and 'shar'. 'iid' uniquely identifies each participant in the event and is useful for querying unique members. 'gender' helps to query by gender. 'wave' number also helps in querying in what wave an individual participated. The features of interest are: 'attr', physical attractiveness; 'intel', intelligence; 'fun', how fun a person is; 'amb', ambition; and 'shar', shared interests. Each of these features of attractiveness are rated in one of two ways:

1. ratings based off a scale ranging from 1 to 10.
2. daters are asked to distribute 100pts to the six features of attraction in the manner he or she feels is most important.

Other things to note: features of attraction with extensions "1\_1" relate to the question asked at some point during the event; the first number pertains to the question number and the second number refers to the time of the event (e.g. 'attr1\_1' is question 1 asked about 'attr' at the start, '1', of the event). Likewise "1\_s" refers to question '1' asked halfway, 's', through the event. "1\_2" refers to question '1' asked the day after, '2', the event. "1\_3" refers to question '1' asked three weeks, '3', after the event. These extensions are important to see, if after various dates, daters change their outlook on stated preferences of attractions with respect to the beginning, middle, day after, or three weeks after the dating event.

There were various things to do with inputs of the dataset. To illustrate a subset of the features space, here is what a sample looks like:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| iid | gender | wave | attr1\_1 | intel1\_1 | sinc1\_1 | fun1\_1 | amb1\_1 | shar1\_1 |
| 20 | 1 | 1 | 100 | 0 | 0 | 0 | 0 | 0 |

As noted previously, 1\_1 denotes to question 1 asked at the start of the event. The Dataset Key that accompanies the dataset states that question 1 is what participants look for in another person. The sample here is a male who put all the weight towards physical attractiveness. Of all the participants in the dataset, only three people put over 90% weight on physical attractiveness: 2 males and 1 female.

Things to do: 1) select features with highest number of samples available. 2) exclude samples with NaN values. 3) Typecast fields like income, SAT scores, and zipcode.

Questions to ask yourself when writing this section:

* *If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?*
* *If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?*
* *If a dataset is****not****present for this problem, has discussion been made about the input space or input data for your problem?*
* *Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)*

**Exploratory Visualization**

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

* *Have you visualized a relevant characteristic or feature about the dataset or input data?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

**Algorithms and Techniques**

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

* *Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*
* *Are the techniques to be used thoroughly discussed and justified?*
* *Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

**Benchmark**

The benchmark model that my project will be compared against is a compilation of two linear models. Fisman and Iyengar proposed several models to determine decisions of participants based off what daters believed important. In fact, looking through the attached article. "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment," shows that the two authors created new features by using old features and defined a decision variable as a weighed mean of new features. The formulae that will be investigated from the author's papers are:

1. the Decision function dependent on attractiveness, ambition, and intelligence.
2. the Decision function dependent on new features formed from ratings received from a dater's partner and self rating of the dater.

Finally, the author summarized results for both these Decision function in tables with various statistics: AVG, STD, R2, and so on. The assumption that the authors made in their paper is that people when dating have "straightforward behavior," meaning that a person dates someone he or she likes more.

Questions to ask yourself when writing this section:

* *Has some result or value been provided that acts as a benchmark for measuring performance?*
* *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

**III. Methodology**

*(approx. 3-5 pages)*

**Data Preprocessing**

As mentioned in the solution section, this project will start off by taking subsets of the dataset and perform data clean up. A quick look at the data set reveals that some of the features do not fall within the specified ranges laid out in the Key file. An example includes entry fields based on the Likert scale from 1 to 10. Some fields have violated these ranges and need to be fixed to conform to the range. Numbers below 1 will be raised to 1; number greater than 10 will be lowered to 10.

Another thing to do with Likert Scale values is to convert them to a scale of 0 - 100 and make sure that values of attraction that are being converted add up 100. To accomplish this I added up the scores on the Likert scale and divided each feature of attraction used by the sum. Afterwards, I multiplied by a hundred. This produced the desired range to compare with other ratings given through the dating events.

The dataset had other issues. The instance in mind is where daters were asked to distribute 100 points across each features of attraction in question. Some respondents accidentally distributed more than 100 points across the six features of attraction. This is no problem because these distributions are relative to how the these points were assigned. Normalizing these responses will rescale everything to lie between 0 and 1. Once this subset is cleaned, I will use univariate analysis to see how features of attraction evolve over time.

The second task at hand will be to create a model that will learn from the dataset and attempt to make matches. What I want to explore in this section is if people clustered by sex match with people in the same cluster for the opposite sex. An example would be a male who lands in the 'clubbing' cluster matches with a female from the 'clubbing' cluster. A scatter matrix plot will help to identify correlation between features. This will help to reduce the number of dimensions and, perhaps, reduce the problem so that the clustering is visually interpretable. The clustering algorithm of choice will be K-means and the number of clusters will be chosen with respect to the silhouette score.

Constructing the matching algorithm will require me to choose what features to include. Choosing the features for this algorithm will depend on the structure of the mentioned clustering above. The dataset includes ratings that dates give each other and will be incorporated in the algorithm. At the moment, I will need to start the project and allow creativity to dictate the direction of the project.

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

* *If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?*
* *Based on the****Data Exploration****section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?*
* *If no preprocessing is needed, has it been made clear why?*

**Implementation**

To verify that men value physical attraction, I will be taking a subset of the original dataset. The subset in mind identifies each person uniquely by 'iid'. Based off this subset, my approach will be to make two more subsets: one for males and another for females. At this point, data will need to cleaned in the sense that missing data points will need to be filled, dropped, or corrected if not within the Data Key file's allowed range for each feature. Once cleaning is done, I will normalize the features of attraction being investigated. The features of attraction will be divided into four categories:

1. how individuals rated relevant features of attraction at the start of the event.
2. how individuals rated relevant features of attraction halfway through the event.
3. how individuals rated relevant features of attraction a day after the event.
4. how individuals rated relevant features of attraction three weeks after the event.

Afterwards, a histogram will be created to investigate how preferences of attractions change over the course of the speed dating event. Basic statistics, or frequency charts, will aid in deciding whether there is an overwhelming preference for a feature.

The clustering aspect of the visualization will use the aforementioned subsets formed from data cleansing. I will be using K-means clustering with a Euclidean metric to determine distances between points. My clustering algorithm will use several cluster values to figure out the number of optimal centroids.

Once visualization is completed, I will develop a model to determine matches. The visualization will help in deciding what features to use in my model. Moreover, other choices to consider for feature selection are forward or backward feature selection. Finally, linear models that incorporate feature selection are Ridge Regression or Lasso algorithm. The learning algorithm will depend on the structure of the dataset once data exploration is completed.

Questions to ask yourself when writing this section:

* *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
* *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
* *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

**Refinement**

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

* *Has an initial solution been found and clearly reported?*
* *Is the process of improvement clearly documented, such as what techniques were used?*
* *Are intermediate and final solutions clearly reported as the process is improved?*

**IV. Results**

*(approx. 2-3 pages)*

**Model Evaluation and Validation**

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

* *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*
* *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*
* *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*
* *Can results found from the model be trusted?*

**Justification**

A benchmark for my PCA chosen features with Nearest Neighboors Classifier is not available, but the visualization produced by projecting the most common features produced by ExtraDecisionTrees, RandomForestTrees, and the f-classifier offers an interesting narrative. Each of these features selection algorithms returned the same top 8 features respectively. These features are 'like', 'attr', 'intel', 'sinc', 'fund', 'amb', 'shar', and 'prob'.

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

* *Are the final results found stronger than the benchmark result reported earlier?*
* *Have you thoroughly analyzed and discussed the final solution?*
* *Is the final solution significant enough to have solved the problem?*

**V. Conclusion**

*(approx. 1-2 pages)*

**Free-Form Visualization**

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

* *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

**Reflection**

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

* *Have you thoroughly summarized the entire process you used for this project?*
* *Were there any interesting aspects of the project?*
* *Were there any difficult aspects of the project?*
* *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

**Improvement**

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

* *Are there further improvements that could be made on the algorithms or techniques you used in this project?*
* *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
* *If you used your final solution as the new benchmark, do you think an even better solution exists?*

**Before submitting, ask yourself. . .**

* Does the project report you’ve written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your analysis, methods, and results?
* Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?
* Is the code that implements your solution easily readable and properly commented?
* Does the code execute without error and produce results similar to those reported?

What influences love at first sight? (Or, at least, love in the first four minutes?)

Data Exploration Ideas

What are the least desirable attributes in a male partner? Does this differ for female partners?

How important do people think attractiveness is in potential mate selection vs. its real impact?

Are shared interests more important than a shared racial background?

Can people accurately predict their own perceived value in the dating market?

In terms of getting a second date, is it better to be someone's first speed date of the night or their last?