**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 [1], 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 ambition and intelligence than men do [1]. 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.

The first thing to do is look for structure in the dataset. This will require cleaning data and working with samples that have the most information possible, that is work with the least amount of NaN values. Visualization will aid in this task, especially if the number of features are reduced to a three dimensional space or lower. Once this is accomplished, choosing a learning algorithm will depend on what data exploration yields.

**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. In the dating context, precision allows for a measure of the correct 'yes' decisions produced by the model. This measurement is a ratio of correct positive 'yes' to the total positive 'yes' (total means the sum of True 'yes' and False 'yes' labeling). Recall is important to the model because it measures the ratio of correct True 'yes' to the sum of correct True 'yes' and correct False 'no'. f1score is a harmonic mean of Precision and Recall and reports accuracy for positive 'yes' responses produced by the learning model.

At first, my evaluation metric was Mean Absolute Error (MAE). The goal was to benchmark with results published by Fisman and Iyengar. However, I realized that the authors of the article incorrectly used a regression algorithm for classification. I have no way to compare the weights for a hypothetical linear classification model. However, in spite of that, the two professors did publish basic statistics for the features that they selected. This will aid in doing sanity checks in lieu of the model chosen by the authors.

**II. Analysis**

**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 helps in querying 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 Likert 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 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, with 'iid'= 20, who put all weight towards physical attractiveness (this participant distributed all 100 pts to physical attractiveness, 'attr'). Of all the participants in the dataset, only three people put over 90% weight on physical attractiveness: 2 males and 1 female.

Quick things to note in this step: ratings were changed from a Likert scale of 1 - 10 to a range from 0 - 100. This was done prior to computing basic statistics on the dataset. Prior to scale changing, most ratings were on a scale of 0 - 100. With that on the record, the mean weight that males report for physical attraction in a mate is 26.97 (STD 13.39) as opposed to the mean weight of 18.02 (STD 9.93) for females. The difference between what each gender is looking for at the start of the event is 8.95.

Likewise, comparing what males look for regarding intelligence, 19.43 (STD 6.81), against what females look for at 20.99 (STD 6.82) are values close to each other; STD is almost equal as well. At the exploratory stage, it is not apparent that women put more weight on intelligence than men do. Later steps will show this. However, there is disparity in ambitiousness sought by gender. Males, on average, weighted ambitiousness with 8.78 (STD 5.93) and females weighted ambitiousness at 12.82 (STD 5.63). From this exploration, ambitiousness is something that females weigh more than males do. The ipynb file has more statistics to pour over and the reader is invited to go explore the file.

Other things apparent from skimming through the dataset is that there are lots of NaN values once loaded into a Pandas DataFrame. Part of this is by design of the experiment; some fields were left empty as a control for the experiment. In addition, the Data Key file states that funding was an issue in gathering samples. Other explanations include some fields being left empty by respondents accidently, discomfort caused by the question, or that the question was not applicable to the person. Finally, outlier detection by Turkey's Method yielded 239 samples across 15 different features.

**Exploratory Visualization**

Below are histograms of what people are looking for in terms of physical attraction. The vertical represents number of respondents; the horizontal represents weight given.



These plots are intended to show whether or not there is a shift of preference over time. In this case, at the beginning of the event, a day after, and three weeks after the event. The blue bars represent what people are looking for at the start of the event. The green bars represent what people are looking for a day after the event, and the red bars represent what people are looking for three weeks after the event.

Disclaimer, the histogram used a setting that 'stacked' the bars on top of each other. Plotting them without 'stacked' yielded messy visuals. The way to interpret the visuals above is to look at the heights of each bar relative to its color. For example, the blue bar for 'attr1\_1' for males peaks at 20 with a frequency of 125 respondents. Likewise, the green bar for the same plot has about 100 respondents and peaks at 20. Finally, the red bar in this plot also has its peak at 20 with about 50 respondents. The mean for male respondents for this features is 26.98 (STD 13.39).

The histograms also obviates that females do not weigh physical attractiveness as heavily as males. For example, attractiveness was between 10 and 20 as opposed to males who replied with numbers centered near 20. Female response rate also dissipated starting with 150 replies, falling to 100, and then to about 75. The height of bars over time is consistent relative to each color. Blue, green, and red bars have max frequency at 10 - 20 weight. The average for this feature with female respondents is 18.02 (STD 9.93).

More histogram plots are included for features in the ipynb file. Again, the reader is invited to consult these plots. Something powerful from having these frequency plots is that they are a visualization of the distribution for a feature. Both genders have a positively skewed graph for physical attraction. This helps in computing the probability of extreme values of the three people who weighed physical attractiveness at over 90 percent.

**Algorithms and Techniques**

Previous work in this domain is inspired by Chris McKinlay's analysis on OkCupid data [5]. McKinlay used boosting on categorical data to figure out how to weigh features to the website's matching questionnaire. The main finding that resulted from using a clustering algorithm is his visualization. McKinlay found categories into which females were lumped and graphed these on a three dimensional axis.

What is dissimilar with McKinlay's dataset and this one is that the dataset for this project has both categorical and numerical responses. I'll hypothesize there exists some inherent structure and use McKinlay's finding to create a visualization.

A Linear Classifier will be used to stay true to the spirit of the authors who compiled this Speed Dating Dataset. However, I am skeptical about how the authors used linear regression in a classification setting.

For the Linear Classifier, I will implement GridSearch on a Support Vector Classifier (SVC) from sklearn. A list of things that GridSearch will do for the Linear Classifier are listed below:

* Try different values of 'C'. 'C' is the separating distance, or margin width, between 'yes' and 'no' decisions that are to be predicted. The figure on the right illustrates the green separating line and the black margins.
* Do cross validation with 10 bins. Cross Validation will train on 75% of available samples and test on 25% of the remaining available samples. These sample are ones described in an earlier section, that is, samples with the least amount of NaN values.
* Use an f1\_scorer to measure accuracy.

Intuition helps in thinking about how the Linear Classifier works. For instance, in a 2-D, or 3-D, feature space, samples are plotted. If the feature space is 2-D, a line is the separating boundary as shown in the figure above. For a 3-D feature, a plane is the separating boundary between decisions. Higher dimensional settings use higher dimensional hyperplanes. The biggest assumption that the authors who compiled the Speed Dataset make, without stating, is assuming that the space is linearly separable.

**Benchmark**

The benchmark model provided with the article is flawed. Fisman and Iyengar proposed a Linear Regression Model to determine decisions of participants based off what daters believed important. Their Decision function is dependent on attractiveness, ambition, and intelligence. Below is a table summarizing the authors results:



For this project columns (1), (2), and (3) are explored. Columns (4), (5), and (6) are not explored at this time because the analysis of these columns adds to the features space rather than simplify it. Interpreting Columns (1), (2), and (3) is an easy feat. Column (1) corresponds to females, Column (2) corresponds to males, and Column (3) corresponds to both genders taken together in the Regression Model.

The model that the authors of the study used is below:

key_for_equation.PNG

The row with attractiveness gives the coefficient for females to be 0.119 and for males 0.140. Males put a greater weight on attractiveness by 0.021. Both genders lumped together in the model yielded 0.119 weight for the attractiveness features.

In a similar fashion, the coefficient for the Linear Model for females gave a coefficient of 0.045 as opposed to males who ended up with a coefficient of 0.023 (a difference of 0.022). Females weigh intelligence more than males do.

The Linear model put ambitiousness at a tie, but its relevance is indisputable. The authors of the study point out that females seek ambitious males because of the ability to acquire resources; males care about ambition because males do not want to date a woman more ambitious than he [1].

Finally, the R2 values help in seeing whether a model is over fitted. By quick inspection, one would think that the middling R2 values for "Linear Regression" is low and no better than chance. The problem with the dataset is that the data is noisy. It will be shown later what reducing the features space helps in visualization what sort of model to apply. It is important to keep in mind that the metric of choice for my model is f1\_score and will be compared with R2 from the authors as a benchmark.

**III. Methodology**

**Data Preprocessing**

The first thing I did before doing any preprocessing was figure out with what I was working. I ran a function I defined in a separate file called features\_creator.py. The method I implemented counts the number of samples for each feature. I let my threshold of samples be 7302; the list called features\_space defined in features\_creator.py file includes features with the number of samples mentioned above. Features whose sample size fell below that threshold were not considered for this project.

As mentioned in the solution section, this project will start 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 by the sum. Afterwards, I multiplied by a hundred. This produced the desired range to compare with other ratings given to a date during the event.

The dataset had other issues. The instance in mind is where daters were asked to distribute 100 points across each feature of attraction in question. Some respondents accidentally distributed more than 100 points across the six features of attraction. This was not a problem because these distributions are relative to how the these points were assigned.

After the features were scaled to 0 - 100, I ran the dating\_attributes\_vs\_time\_describe() function defined in the features\_creator file. This method takes in a DataFrame and gender as input and displays the min, max, mean, and STD. I also used the dating\_attributes\_vs\_time\_hist() defined in the features\_creator file. This method utilized the plot() function with parameters kind = 'hist', stacked = True, bins = 10. 'hist' tells plot() to generate a histogram with the setting 'stacked' set to True. 'stacked' places the bars in a histogram on top of other bars that fall in each respective bin. A bin is one bar along the horizontal. The histograms seen in the Exploratory Visualization section are outputs from the mentioned function.

Points of contention included categorical labels that needed to be changed to numerical values. Examples are SAT scores, zipcode, and tuition. These needed to be changed to numerical values for feature selection by tree method. Moreover, for PCA I needed to scale the range of the features being used to positive numbers between 0 and 1.

**Implementation**

The best visualization that shows relationships between some of the features is the correlation scatter matrix shown below (zoom in to see).



When I went through basic statistics, I noticed that when one feature went up and other features values went down. Darker shades indicate positive correlation one way or the other. Specifically, green indicates positive correlation; blue indicates negative correlation. Having this graph helped me in deciding if Principal Component Analysis is useful later in my analysis. The goal is to reduce the number of features to simplify my learning model.

There is some insightful information looking at the graph above. Caveat: the graph produced above is for both genders. The reader is invited to consult the ipynb file that accompanies this report and consult the correlation matrices for males and females respectively.

The features chosen for the correlation were specifically picked out of all the available features because these features had the most samples. The graph above was produced after data clean up.

That said, one feature that caught my eye was that shar2\_1 is correlated with shar1\_1. This result is amusing because 'shar2\_2' pertains to the question of what a dater believes the person of the opposite sex is looking for in regards to shared interests. 'shar1\_1' pertains to what the mentioned person is looking for in shared values. One interpretation is that a person can change the weight for what is sought based off what is believed to in demand. Another way to read things is that when dating, you are 'vibing' off each other, gauging the situation, and seeing what the odds of a second date are.

Self-perception also has its affect on what is pursued. For example, 'sinc3\_1' is correlated to 'sinc1\_1'. 'sinc3\_1' pertains to the question of a how a person rates himself, or herself, for that feature. In plain English, this means that a sincere person looks for a sincere person. The reason why this is insightful is because of literature in the field [2]. Nate Silver, the famous statistician, wrote an article that states people look for features that they themselves have. Some correlated features fall in line with Silver's article.

What stands out the most is 'attr2\_1' effect on 'attr1\_1' and vice-versa. This correlation implies that a person's belief on what people look for in terms of attraction affects what he, or she, wants in the opposite sex. The OkCupid blog had an interesting article on the matter. The article highlights that people lie about themselves online. Moreover, males tend to pursue the unachievable, and females tend to drop their standard when immersing in online dating [3]. In fact, another work also holds that females tend to be more attractive on average than males [4]. Dating Literature helps guide that the project is heading in the right direction.

Finally, just by exploring the correlation matrix, my hypothesis is that having information is key as to how to speed date. If a dater knows, or has some insight as to what the date wants, he or she can be successful in the sense of setting expectations. I did not discuss negative correlation, but the example to note on that would be 'attr1\_1' is inversely correlated with 'sinc1\_1', 'intel1\_1', 'amb1\_1', and 'shar1\_1'. At face value, attractiveness takes weight from everything else. Thus far, results from this visualization fall in line with published results.

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?*

Works Cited

[1] Fisman, Ray. Iyengar, Sheena. "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment."

[2] Silver, Nate. "In the End People May Really Just Want to Date Themselves". https://fivethirtyeight.com/features/in-the-end-people-may-really-just-want-to-date-themselves/

[3]"Your Looks and your Inbox". OkCupid Blog. https://theblog.okcupid.com/your-looks-and-your-inbox-8715c0f1561e

[4] "Why Beatiful People Have More Daughters: From Dating, Shopping, and Praying to Going to War and Becoming a Millionaire." Miller, Alan S. Kanazawa, Satoshi.

[5] "How a Math Genius Hacked OkCupid to Find True Love". Wired. https://www.wired.com/2014/01/how-to-hack-okcupid/