**Project Problem and Hypothesis**

* What's the project about? What problem are you solving?
  + The project is about seeing if there is a relationship between an equity investment fund’s performance and if it’s heavily invested in stocks that are widely covered by the Street (investment brokers). I will be using 7-year performance.
  + I want to know if funds that invest in a lot of stocks that hare highly covered by brokers tend to underperform their benchmark. It could be a useful tool when making investment decisions.
  + For simplicity purposes, I want to look at funds that are benchmarked against the S&P 500, the Russell 2000, MSCI EAFE, and the MSCI Emerging Markets indices. I would assume the results would differ for funds that invest in each asset class.
* Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?
  + I hope to predict a new feature called “broker score”. The broker score is the sum of all the fund’s investments’ broker coverage total. If the fund invests in 2 stocks that have a broker coverage of 20 and 33 respectively, then that fund’s broker score would be 53. I want to predict a broker score for each asset class and if a fund’s broker score is greater, it has a significant chance of underperforming its benchmark.
* What kind of impact do you think it could have?
  + I believe is could aid investors in selecting the investments for their portfolio.
* What do you think will have the most impact in predicting the value you are interested in solving for?
  + The broker coverage of each stock.

**Datasets**

* Description of data set available, at the field level (see table)
  + Universe will be Mutual funds with an inception date prior to 12/31/2005.
  + 7 year performance starting at 12/31/2005.
  + Mutual funds “broker score”, which is the sum of each of its holdings broker coverage total. This is data that I have to collect and crunch.
  + 7 year performance of the S&P500, Russell 2000, MSCI EAFE, MSCI Emerging Markets.
* If from an API, include a sample return (this is usually included in API documentation!) (if doing this in markdown, use the javacription code tag)

**Domain knowledge**

* What experience do you already have around this area?
  + In my current profession, I do conduct manager diligence on investments for a wealth management firm.
* Does it relate or help inform the project in any way?
  + Yes. It could impact how I view certain investment managers and be another metric I use when viewing each asset class.
* What other research efforts exist?
  + I am looking to quantify the efficient market hypothesis and apply it directly to investment funds. The efficient market hypothesis is an investment theory that states it is impossible to "beat the market" because stock market efficiency causes existing share prices to always incorporate and reflect all relevant information. I am assuming that the more broker coverage a stock has the efficient it is.
  + There is a lot of research on the theory of the efficient market hypothesis, but I would like to directly apply it to the real world application.

**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 am concerned with what the best way of structuring the data.
* What are the assumptions and caveats to the problem?
  + I using data at 12/31/2005. Ideally I would like to do it on a rolling basis rather than at a certain point in time. But the collection of all that data would be extremely difficult and tedious. I am using 12/31/2005 and the 7 year performance from that date, because there was a full market cycle during those 7 years.
  + What data do you not have access to but wish you had?
    - This data on a rolling basis. Having data every year for 10 years and taking the average would probably be more accurate.
  + 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)
    - It is not representative of the whole sample size. I am only taking into account stocks that are held in the benchmark. If a fund invests in a stock not held in the benchmark, it will not be capture. Pulling this data would be too arduous and time consuming.
* 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?)
    - If the modeling ends up being wrong, it could lead to making bad investments in certain investment funds.
    - If the modeling is correct, it would just be another tool to use in analyzing the pros and cons of making an investment in certain funds.
  + Is any of the data incorrect? Could it be incorrect?
    - The dataset I am using has survivorship bias. If a fund was a poor performer and closed, it will not be included in the dataset. This is something to seriously consider and note.

**Outcomes**

* What do you expect the output to look like?
  + I expect the output to show that there is a certain “broker score” level for each asset class where if an investment fund is below that number the fund will outperform its benchmark. Maybe the further below the “broker score”, the more it outperforms.
* What does your target audience expect the output to look like?
  + The target audience would like to see the output show a broker score per asset class. Funds that have a broker score below that number have a statistically significant chance of outperforming the benchmark.
* What gain do you expect from your most important feature on its own?
  + The most important feature would be the “broker score” and 7 year performance of the fund and benchamrk
* How complicated does your model have to be?
  + The model should not be that complicated. It would be looking at the broker score and the performance relative to its benchmark.
* How successful does your project have to be in order to be considered a "success"?
  + Success would be defined as there is a specific broker score exists, where funds below that number have a significant chance of outperforming. But if there isn’t one, then the hypothesis has no importance, which I would consider a success as it is a feature you do not have to consider when analyzing investment funds.
* What will you do if the project is a bust (this happens! but it shouldn't here)?
  + If the project is a bust that means the hypothesis of investment funds that invest in companies with a lot of broker coverage does not have any impact on relative performance and should not be considered heavily when making investment management decisions.