New York Yankees Statistical Analysis Questionnaire

Michael Hirsch

January 26, 2015

1. What perks do you expect from working for the Yankees?

   I would expect employee benefits including health and commuter benefits from any organization that employees me. With regards to perks that are specific to the Yankees, I would expect to receive discounts and possibly advanced sales on tickets as well as discounts on merchandise.

2. Which statistics do you use to evaluate pitchers?

   To answer this question, I'm just going to consider starting pitchers, since I would evaluate relievers a bit differently. There isn't a single statistic that I think sums up a starting pitcher's skill perfectly, so I like to have a look at the following stats when evaluation a picther's performance:

   - K/9 and BB/9: These straightforward stats are useful because they indicates directly how well a pitcher can locate the strikezone as well as how "nasty" his pitches are. High strikeout rates are appealing because it shows that the picther can use his repoirtoire effectively, and need not rely on his defense retire batters. High walk rates are indicative of an inability to locate the strikezone, and show a pitcher's ineffectiveness in keeping men off base. I think that both of these stats are representative of a pitcher's contribution to his team winning games.
   - GB%, LD%, FB%: These stats are interesting because it allows one to determine what type of picther is on the mound. High ground ball and fly ball rates are attractive, since these type of batted balls are most often played for an out. High line drive rates are troublesome, since on average, these types of batted balls fall in for hits. Also, managers could find these satistics useful in creating a lineup (in the case of the opposing team) or choosing defensive formations (for the pitcher's team).
   - FIP: I think that FIP is a fair measure of a pitcher's performance independent of the defense behind him. ERA is too dependent on the defense and fluctuates more wildly each season than FIP.

3. What are your favorite baseball websites or podcasts? What do you like about them?

   For statistics and analysis, I frequent both Fangraphs and Baseball-Reference. I think that Fangraphs has an excellent interpretation of the game from a sabrmetric perspective. I also am a big fan of visual tools for analysis, which Fangraphs provides for nearly every statistic in the game. They also have very thorough documentation on each sabrmetric stastistic, which is quite useful in conducting independent analysis. Baseball-reference is great for looking up player and team statistics, as they have very thorough records of nearly everything one would be interested in: standard and sabrmetric batting/picthing/fielding statistics, salaries, awards, as well as interesting "similarity scores" that attempt to compare a player to similar players in history. Baseball-reference also has great tools for exporting data, which is very useful in conducting my own analysis.

   For news, I like Bleacher Report's MLB section. They have interesting original analysis, updates on rumors and devloping stories, as well as a thorough compilation of external news sources including ESPN, Pinstrip Alley (in the case for the Yankees), MLB originals, and more. Bleacher Report also

has a great app for iOS, allowing me to keep track of specific teams and players, devliering notifications when important news breaks.

4. Which of this off-seasons free agents (unsigned or already signed) do you think will provide the best value?

   There are a few contracts that jumped out at me this offseason.

   - Chase Headley: Chase's solid defense and decent bat will prove to be a great pickup for the Yankees. His deal was substantially smaller than Pablo Sandoval's deal with the Red Sox, and I think that Chase is a superior player in many respects. Headley is getting $\$52mil/4yrs =$ \$13mil a year while Sandoval is receiving $\$95mil/5yrs =$ \$19mil a year, meaning the Yankees are getting a quantitatively similar player for \$6mill less a year. Headley will also provide a level of defensive at third that the Yankees have not seen in years, as he ranks among the best among third basemen in defensive runs saved. His bat is not stellar, but I think he will be more productive than A-Rod can be after missing a year because of suspension and nearly an entire season due to injuries in 2013.
   - Andrew Miller: I am really excited about having Andrew Miller team up with Dellin Betances in the back end of the Yankee's bullpen. While it was sad to see Robertson go, I think signing Andrew Miller was important for a few respects. Firstly, resigning Robertson would have sacrificed the Yankees a draft pick, a resource that I don't think the Yankees should give up. Secondly, Miller was on par pitcher with Robertson in many respects and the Yankees signed him for less money than it would have cost to resign Robertson. In fact, Miller bested Robertson in WHIP, K/9, BB/9, and FIP. If Miller can pick up where his stellar 2014 season ended, the Yankees will have a solid righty/lefty combination to close out games.
   - Stephen Drew: While Drew had an absolutely horrific 2014 season, his 1 year \$5mil contract is a great deal for the Yankees. Since trading Martin Prado to the Marlins, the Yankees were in need of a second baseman. While Refsnyder appears to be a solid prospect at second, it is important for the Yankees to have an established player at the position while Refsnyder develops into a major leaguer, especially with a very young Didi Gregorious at short. Drew's defense is solid and did not suffer last year, so as long as he can fix his offensive struggles, this will prove to be a solid pickup at a great value.

5. With data that tracks the location and movement of every fielder and baserunner when the ball is in play, what questions would you want to research, besides measuring a fielders range?

   With ballparks now rolling out technology to track this kind of information, I am really excited to see what new kinds of statistics will be introduced. I see the following ideas as ones I would be interested in researching.

   - Baserunning Leads: Being able to track the size of a baserunner's lead could provide extremely valuable statistics for player's stealing ability, a catcher's ability to throw out runners, and a pitcher's ineffectiveness in keeping runners close. Analysis could identify a maximum lead size for each pitcher in which a stolen base would be successful. Also, pitchers and their coaches could keep track of when a baserunner has exceeded his range in which he can successfully return to a base, allowing for more knowledgable and data driven throws back to the base, increasing the chances of a pickoff.

- Baserunner's Movement: Aside from looking at leads, I think it would be imortant to evaluate baserunners' acceleration and deceleration, especially when going for extra bases. While there is a multitude of information that determines whether a runner should run home from third, I think that it would be interesting to predict whether a runner who stopped at third could have actually made it home given his speed and acceleration. This would provide a new measure amounting to something like "Forgone Runs," which would be an indicator of a players (and potentially the third base coach's) ability to properly send a run home.
- Optimal Placement for a Cutoff Man: A good cutoff position and throw is vital in saving runs as well as preventing players from getting extra bases on a hit. Measuring the shortest path from a ball in the outfield to a base through a cutoff man could prove to vital for defensive formations and placement of the cutoff man.

6. Suppose you are asked to build a model that predicts salaries for arbitration-eligible players and are given a list of possibly-relevant variables. Walk through your steps for this research. How would you test your model?

   Predicting salaries is a regression problem, so we will need to use a model that predicts continuous values, rather than categorical. I am a fan of the random forest technique, which is an ensemble method that creates multiple regression trees on random subsets of the training data and predicts a value equal to the average of the predicted values of the individual trees. One benefit of such an approach is that it can aid in feature selection, helping to determine which of the variables are important.

   (a) Acquire data about past arbitration contracts containing all of the relevant variables, as well as the size (millions) and length of the contract (years).

   (b) Create a computed variable that indicates the salary per year of contract (millions/years).

   (c) Conduct preliminary analysis of the data. Plot salaries against the variables to try and get a feel for what variables might be most important.

   (d) Split dataset into a training and a test set by taking random samples.

   (e) Train the random forest model using the training set. Parameters to be tested inclue the number of variables used at at each level in the regression trees as well as how many trees should be built.

   (f) Asses the accuracy of our model by predicting salary values for the test data and comparing these to the actual salaries. Plot the size of our forest against the mean square errors to what the best forest size is.

   (g) Use the model to predict salaries for arbitration eligible players.

7. What statistical techniques do you know that have applications in baseball research? What questions in baseball do you want to study that would take advantage of these skills?

   Techniques that I am aware of that have use in baseball research include regression analysis, Monte Carlo simulations, sabrmetrics, and classification. While I find all things baseball very interesting, I would have to say that I am most interested in studying what happens in preparation for and within a particular game. Optimal lineup creation, defensive alignments for particular batters, and pitching sequence prediction can all be studied using simulations. As mentioned earlier, making use of new

tracking technology to design new metrics that can better classify every aspect of a player can be done using classification and regression techniques. Apart from things occurring with respect to a particular game, I would also be interested in scouting, both nationally and internationally. Having a talented farm system as well as making smart trades and free agent acquisitions is very important to the quality of a team's major league club and therefore I think that it is important to invest a fair amount of research time into recruiting top talent. Also, given the new relations that the United States has with Cuba, I see a lot of untapped potential that could be drawn upon using data driven research.

8. Describe a project you worked on (baseball or non-baseball, academic or personal, solo or in a group) that demonstrates the skills you would bring to the Yankees.

   Since 2011, I have been involved in numerous data projects, which can be seen on my resume. Some relevant projects include monetary projections for finance teams, customer interaction projections for sales and marketing teams, schema design in a data model, data translation, and classification of mined data. Currently, I am doing research in statistical learning theory, the mathematical foundations of machine learning, using baseball data to drive my research. Skills that I would bring to the Yankees are ownership of projects, vast experience with data, creativity, and eagerness to learn.

   I think that projects I have worked on at Booker and Relationship Science could speak to my ability to work well in an office setting as a data engineer, but I think that my current research work is most relevant to my role as a statistical analyst for the Yankees. I am currently writing a research paper on a technique from machine learning called random forests using data about batters in attempts to classify a hall of fame candidate. This is entirely independent research that originated in my interest in 2015's hall of fame class. As far as I am aware, no such model has been created in the past, and I think this concept is representative of my creativity and my eagerness to learn. I have spent numerous hours teaching myself the requisite knowledge for such an undertaking, consulting textbooks, online material, and examining statistical packages in R in order to better understand the techniques. Also, I have been able to demonstrate my skills with data and mathematics by applying to them techniques that I have learned.

   What excites me most about working at the Yankees is the opportunity to work with new data that has a direct and intrinsic meaning to me and baseball fans in general. I am also eager to help the organization make data driven decision in all aspects of baseball operations. Thoroughly learning the techniques used in statistical analysis for baseball and familiarizing myself with the metrics that the organization uses will be a mutually beneficial endeavor. I think the future of baseball will be heavily influenced by statistical analysis, and I am excited by the prospect of heading there with the Yankees.