ORFE UNDERGRADUATE FUNDING PROPOSAL

Michael Chiang '17
Alan Du '17

Given our shared interest in professional sports and knowledge of Statistics and Machine Learning (SML), Michael Chiang and I (Alan Du) are working on a Sabermetrics project. The goal of the project is to use SML techniques to select the best teams in Daily Fantasy Baseball competitions.

We hope to apply the knowledge and techniques that we have learned in our ORF classes over the last three years:

- **ORF 350 and ORF 405**: Regression techniques for analyzing big data. A single season of Major League Baseball has 2,430 regular season games, and there are 25 players on a team's active roster. This means that for a single season we have a sample size of approximately (2,430 games) * (2 teams/game) * (25 players/team) = 121,500 for the purpose of analyzing player performances. The data will also be high dimensional due to the vast array of baseball statistics and other features that may predict player performance.
- **ORF 363:** Optimization techniques and machine learning algorithms.
- **ORF 574:** Optimal money management (Kelly Criterion) and risk management.

Below we have outlined the five stages of our project. We are currently in Stage I.

Stage I: Data Collection
Before beginning our regression analysis, we must compile a dataset that includes historical game logs and fantasy point contributions for every player in Major League Baseball. We seek game logs that provide a comprehensive range of player statistics so that we can analyze a rich set of features in Stage II. Gathering this historical game data for every player has been challenging (see budget request). We will also consider compiling additional features outside of game log statistics, such as team records, rest days, hot or cold streaks, stadium information, and weather.

The result of Stage I is a comprehensive player-level dataset, split into testing and training sets.

Stage II: Modeling & Forecasting Using SML Techniques
We will apply a feature selection technique to pick a subset of features to predict the fantasy point contributions for any given player. The specificities of the model used to predict fantasy point contributions is to be determined – one natural approach we are considering is Bayesian linear regression. A look at fantasy sports literature suggests that a Bayesian approach may be appropriate. Additional ideas gathered from literature include using probabilistic matrix factorization to model the dependency between two opponents and applying clustering techniques to model players sharing some feature separately. We will use the testing set to evaluate our model.

The result of Stage II is a model that can predict fantasy point contributions for any given player.

Stage III: Combinatorial Optimization
Based on the fantasy point contribution projections from Stage II, we will select the optimal set of ten players for our fantasy team. To do so, we will solve a constrained optimization problem over a feasible set defined by salary cap restrictions. This is a 0-1 multiple knapsack problem

that can be solved using dynamic programming or approximation algorithms. Since the problem is NP-complete, we will need to determine a strategy for eliminating some players to avoid computational complexity issues.

The result of Stage III is an optimal lineup for any given day.

Stage IV: Back Testing
We will check how our lineups would have performed in historical daily fantasy pools. One method to evaluate the performance of our lineup is to check if it ranks high enough in the pool to receive a payout that exceeds the entry fee. Another method is to see if the lineup beats the lineups suggested by "experts" on daily fantasy sports websites. If the performance of our lineups is underwhelming, then we must return to Stages II and III and revise our approach.

The result of Stage IV is a validation of our methods in Stages II and III.

Stage V: Live Testing
We will test if our model can generate lineups that consistently beat "expert" suggestions on daily fantasy sports websites.

Budget:
Stages I-IV:
$799 for historical game log data for every player in Major League Baseball (2007-2015), provided by FantasyData LLC.

Stage V: (contingent on success in Stages I-IV):
$499 for one month of in-season game log data data, provided by FantasyData LLC.

FantasyData LLC pricing information available at:
http://fantasydata.com/pricing/mlb-data-api.aspx#mlb-historical-database

References:

- http://cs229.stanford.edu/proj2015/104_report.pdf
- http://cs229.stanford.edu/proj2012/Kapania-FantasyFootballAndMachineLearning.pdf
- http://hips.seas.harvard.edu/files/adams-dpmf-uai-2010.pdf
- http://blog.smellthedata.com/2011/02/thoughts-on-modeling-basketball.html
- http://datashoptalk.com/double-yo-money/
- https://futureworlds.com/premier-league-picks-machine-learning-fantasy-football/