JQAS Working Manuscript

The Lemanski Sports Analytics Group

April 26, 2014

Task Summary

Manuscript Due: July 15
Current Assignment:

- Lucas - Luck
- Xinran - Popular Methods
- Marcos - Bracket Recommendation (in Recommendations) [ANDY SAYS: Decision Theory section?]
- Yuhyun - Lit Review

1 Introduction

Every March, millions of people take to an American tradition, filling out an NCAA tournament bracket. Typical strategies include listening to so-called experts or following one's intuition, although the rise in sports analytics and the popularity of sites such as Nate Silver's fivethirtyeight.com are increasing the use of data-driven methods. With this work we review analytic approaches to forecasting NCAA tournament games and introduce our modeling framework designed to capture matchup effects. Many common methods for predicting outcomes assess overall team strength using ranking based metrics. In addition to quantify overall team strength, our methodology provides a data-driven approach to assess team by team match ups and address statements of the form: *"Team Y is a tough draw for team X due to their tempo, size, athleticism, three point shooting, ect.."*. The essence of of the matchup effects is to discern the existence and magnitude of team-by-team match ups. While there certainly is a high degree of uncertainty involved in predicting outcomes and a small number of games for evaluation, we demonstrate the efficacy of our matchup effects. We predicted Duke would lose. THE END

1.1 Common Prediction Methods

Many commonly used methods or models include a seeds based approach or using one of several rankings systems: Sagarin, Pomeroy, ESPN BPI, ect..

1.2 Other

We should be explicit about the Kaggle comp. vs. just filling out a bracket. For kaggle, there are no broken brackets.

1.3 LitReview

Literature Review: What people have done for predicting tournaments before (both bball and other sports are fine), March madness in general (how many people watch, how much revenue is it worth). Predicting human performance in sports (it's hard), history of Kaggle.

2 Popular Methods

This section reviews several existing

2.1 Rating Based Methods

Sagarin, Pomeroy, ect...

2.2 Ensemble Methods

2.3 538 methodology

2.4 Wisdom of the Crowds?

Just sagarin (Scotland or Andy), nate silver, consensus (Xinran)+espn (include into sensitivity study with Lucas).

-When comparing models: modeling point spread vs explicit prediction models: logistic vs cart vs linear models. CART is maybe not effective here because there isn't any obvious heterogeneity in the space. Linear models seem to work well.

3 Data

For many traditional bracket competitions predictions only need to be a binary result (i.e. win or lose); however, there are competitions for which probabilistic predictions are required. Furthermore, this provides a sensible framework for evaluating various loss functions and computing risk for various prediction schemes. Hence, we restrict our focus to methods that return a probability of a team winning any matchup.

3.1 Common Ratings Components

wins, losses, strength of schedule...

3.2 Other influential factors

home court, height, ect...

3.3 team level data vs. player level data

4 Modeling

One interesting dilemma involves whether the outcomes should be modeled in a binary sense (win or loss) or rather should a continuous metric such as the point spread be used. In theory, point spread provides a means for eliciting the relative strength of one team, although as any basketball fan can attest to the final score is often not indicative of how close the game was. A common strategy is for the trailing team to foul in the closing moments of the game, which can often result in a two point deficit turning into a ten point loss. For this reason, we scrape scores with 2 minutes left in the game [Andy says: I think this can be done]. A comparison of these three data aggregation methods is used on cross-validated data from... [Andy says: hmm... I wonder if there is a way to tune/transform point spread to control for blowouts...]

4.1 Model Specification

The general form for a linear model for point spread follows below:

$$Y_{ij} = X_{ij}\beta + \epsilon_{ijk} \tag{1}$$

where X_{ij} corresponds to the difference in predictors for teams i and j and $\epsilon_{ijk} \sim N(0, \sigma^2)$.

4.2 Calibrating Predictions via predictive intervals

Discuss tuning the tails for predictions

5 Decision Theory - Optimal Strategies

5.1 Probabilistic Forecasting

5.2 proper scoring rules

5.3 Strategy: Min Risk or Max Expected Earnings?

[ANDY SAYS: The entire process could be simulated using the historical probs of seed X vs. seed Y, this would give some range on the ideal (oracle) solution in which the probs are known]

6 Model: Nearest-Neighbor Matchup Effects

When listening to sports broadcasters, claims of the following type are often made *Team X is a tough matchup for team Y due to their ...* . There are two ways to consider this statement: (1) the overall team strength of Team X will be problematic for Team Y or (2) Team X has certain tendencies above and beyond their team strength that will pose difficulties for Team Y. For the first case models of the type Equation 1 will account for the matchup. However, if the second case is present we need a different approach to analytically quantify whatever characteristics may pose difficulties for a given team. This approach is called the Nearest-Neighbor Matchup Effect and provides a means for capturing team level characteristics. For instance, a glance inside the crystal ball would have revealed that Duke might struggle with a team like Mercer due to... The matchup effects is a three step procedure: (1) the typical model as in Equation 1 is fit, (2) for each matchup, past opponents most similar to the current matchup are identified, and (3) an adjustment is introduced that accounts for past performance against similar teams.

6.1 Matchup Effects

The general form for the matchup effects model follows below:

$$Y_{ij} = X_{ij}\beta + \rho(N_i(j)_k - N_j(i)_k) + \epsilon \tag{2}$$

where X_{ij} coresponds to the difference in predictors for teams i and j and $N_i(j)_k$ corresponds to the residual for the k nearest neighbors of $i's$ opponents to team j. In other words, the second term adjusts the expected outcome based on match ups with similar teams. [ANDY SAYS: clean up notation]

6.2 Choosing Neighbors

There are a multitude of ways to select the neighbors. In particular one needs to consider what variables to consider for selecting neighbors, how should those variables be weighted if at all, and how many neighbors should be selected. [ANDY SAYS: Great Application for BAVA: given a set of information the users can identify similar teams which can then important variables can then be identified]

6.3 Tuning ρ

The natural support of ρ would be between zero and one. The interpretation of the extreme points is rather intuitive - with $\rho = 0$ Equation 2 reverts to Equation 1 and with $rho = 1$ the entire residual for similar teams is retained.

7 Evaluation

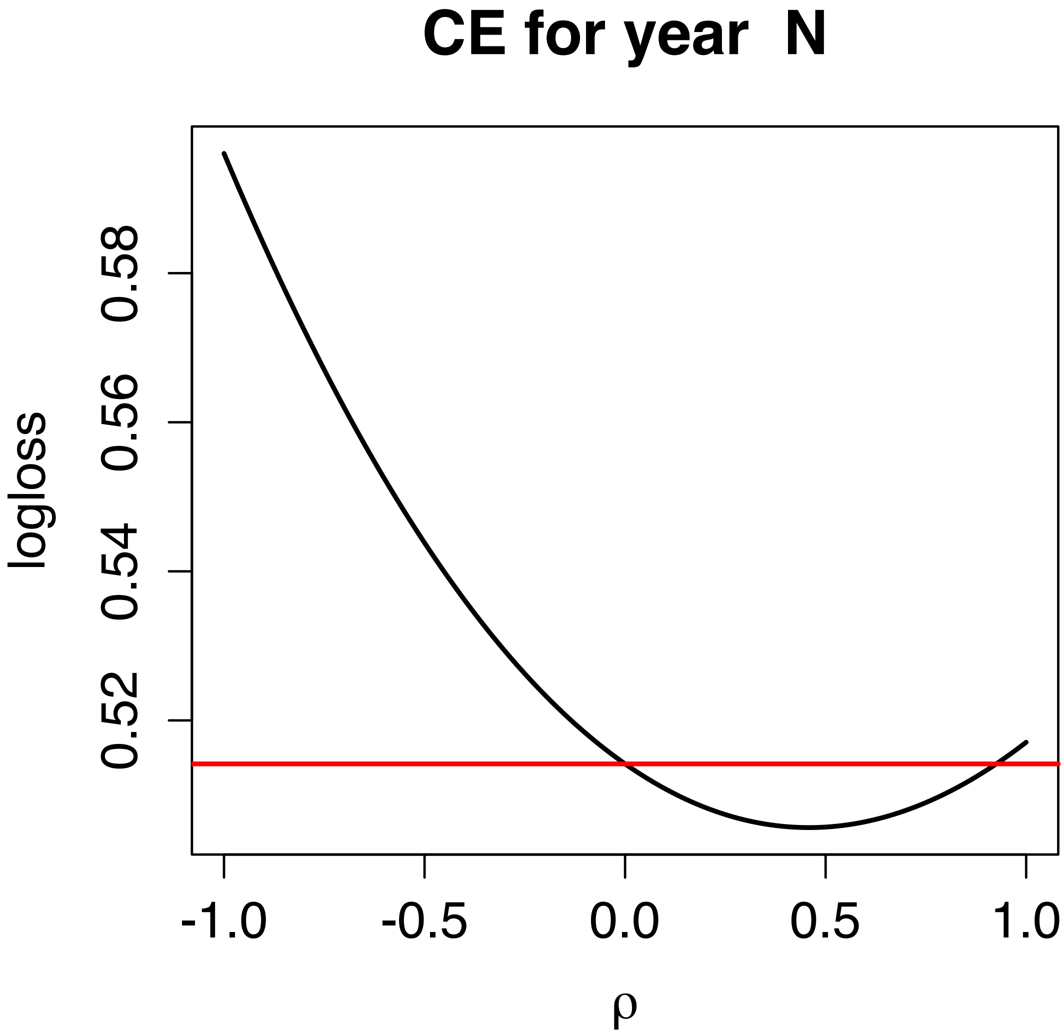

Figure 1: Log loss for range of ρ values

8 Luck

how sensitive is the leader board. It's natural to think 2nd place almost won, but how close was 20th place to winning. (Lucas: sensitivity study)

9 Inference on Transitivity

[Andy says: This may get cut] The transitive property states if $A > B$ and $B > C$ then $A > C$. In terms of basketball consider:

$$P_{A,B} > 0.5 \quad \& \quad P_{B,C} > 0.5 \rightarrow P_{A,C} > 0.5 \tag{3}$$

, where $P_{I,J}$ is the probability of team I defeating team J. Then Equation 3 can be considered a transitive property on basketball match ups. That is if team A is expected to beat team B and team B is expected to beat team C, then team A should also defeat team C. Any sort of rank based approaches would assume this transitive ordering, home court effects nonwithstanding. Note these are probabilities not true outcomes, due to the parity in basketball inferior teams can and often do defeat stronger teams. Nevertheless, our modeling approach can determine if the strengths of a given team present difficulties for a specific team resulting in the transitive property not necessarily holding.

10 Recommendations

things that work and don't work. How do you train the models for each season, do you train the model to predict that years tournament. Or, do you use all 5 years worth of data to predict a single year? The first choice is obviously better. "The data is probably more important than the models- cite from the kaggle winners"

-subsection in recommendations: How should a typical user use this to figure out their bracket? - one strategy is take the most probable. is there another? (Marcos)

11 Discussion

other data such as injury reports (Nate uses this)

Discussion: Is the score at 2 minutes to go better than the final score? In the last minutes of the game, wonky stuff happens

[Andy says: reference place holder] [Gelman and Hill(2006)]

References

[Gelman and Hill(2006)] Gelman, A. and Hill, J. (2006), *Data Analysis Using Regression and Hierarchical/Multilevel Models*, Cambridge University Press.