Applied Mathmamatics 221

Project Milestone 2: Status Update

David Freed and Sam Green

March 29, 2016

Overview of Main Progress

Data Update

Next Steps: Status Update

We have reproduced the next steps that we made in the initial report below, with related commentary for each as to how much progress we have made on each. In sum, we have done almost all of the next steps that we proposed we would do earlier.

Twitter Data Next Steps

1. Get Twitter API keys
2. Identify the best ways to get requests and the structure for streaming data from Twitter
3. Read the literature on sensitivity analyses on Tweets
4. Get trial data (ideally with location tags attached)

Box Score Data Next Steps

1. Create a scrape to get all of the box score data
2. Write a win probability model

The win probability model will likely be a replication of what people already do online and will provide a reference point for how good Twitter is at predicting games. We might initialize the model to the Vegas line or some other ex-ante predictor of performance (Pythagorean estimates are a commonly accepted way of measuring team strength).

Next Steps

Our project proposal demanded that we have three major items in order to run the tests we wanted:

1. Game-by-game Twitter data. As described above, we need this to be able to gauge fan sentiment at every point during the game so that we can get their impression on how the game is going

2. Game data. This comprises both the score at each point in time[1] and the probability of the favorite winning at each point in time

3. Twitter prediction model. For our final results, we will essentially be comparing the standard win probability models above to the Twitter prediction model that we build.

Consequently, we can say that we have accomplished most of our goals. While there are minor things to do on the first two points[2] we are mostly focused now on creating the Twitter model.

To accomplish that, we will need to know how to classify the Tweets by strength of sentiment. Our task after that is to think about how to map those sensitivity analyses to a point spread. Once we do that, we can test our model against basic default models of win probability and see how well Twitter performs, which will give us the desired final results.

[1] Of more importance to us is the difference in scores—whether a team is up 100-90 or 10-0 is immaterial; the ex ante predictors of team strength we used are based in the predicted final score difference.

[2] One thing that jumps out as an immediate next step on the data is matching analog times (i.e. 2:00 EST) from the Tweets to the time left in the game. This will allow us to directly match the predictions by Twitter to predictions made by the win probability model.