ORF387 Final Project

Kenny Huang

kh19

My project idea is to find trends in how NBA players and their relationships affect their decisions when choosing teams in free agency.

I. Abstract

Using a "minibatch" approach, I generated a model that was able to predict whether two players would end up on the same team with (almost) significantly higher accuracy than random chance. Of the variables examined, three-point volume incompatibility appears to have the most notable effect, followed by years played as teammates and age difference. However, when applying these findings to trios of players, no statistical effect from triadic closure could be found. Interestingly, there does exist a noteworthy phenomenon where trios of players in which two pairs have scores above a certain threshold have disproportionately many occurrences of the least close pair ending up on the same team. Although the reasons for this trend are not sufficiently addressed here, it could be an interesting topic to address in future research.

II. Introduction

Originally, I was inspired by the concept of triadic closure and the idea that the relationships between two nodes could indirectly influence their neighbors. In the NBA, every summer, players whose contracts end enter what is known as "free agency", where they are free to renew their contract with their current team or sign a new contract with any other team. In this context, I was curious whether two players who each have a relationship with a third player are more likely to join the same team in free agency than two randomly chosen players. For example, if Player A is "close" with both Players B and C, would Players B and C be more likely to become teammates? To answer this, I first needed to determine what quantifiable factors affect whether two players join the same team, or their "closeness". Once I did so, I could then explore the triadic closure aspect: how players' direct and indirect relationships affect their decisions.

III. Data Collection

After much brainstorming, I came up with the following possible input factors:

- years as NBA teammate
- age difference (integer division by 1000 days or about 3 years)
- same last name
- same jersey number
- same college alma mater
- same international region (USA vs nonUSA)
- 3-pointer volume incompatibility
- free throw volume incompatibility

Luckily for me, I had already had some work done. For a previous personal project, I scraped box score data that actually turned out really useful for this NBA teammate data. I wrote an R script to consolidate my dataset as well as a Python program to summarize it into a table that I could refer to for this project. One unexpected difficulty here was that because I would be evaluating players for each free agency period, I needed to create a separate spreadsheet for the values after every season. For example, Tim Duncan and Tony Parker played 13 seasons together throughout their careers, but if I were to predict Duncan's behavior after the 2005-06 season, I would prefer to know that they only had played together a couple seasons by then, which indicates a weaker bond at the time than near the end of their careers. As such, I ended up creating about 20 very large but sparse spreadsheets.

For the remainder of the player network data, I was extremely fortunate to find two very well-structured and cleaned datasets on Kaggle. From this SQLite database (which I had never used before) and collection of spreadsheets, I was able to extract a lot of data regarding player attributes and performance. In addition, I decided to include additional input variables corresponding to the squares of the teammate years and age differences columns; I don't think the relationship between those two variables and the label is linear, so I wanted to add the flexibility of a quadratic term.

The last thing to collect was the labels. One approach could have been to compile all players entering free agency each year and to find individually where each player ended up signing. However, this would have been very computationally expensive. For simplicity, I decided to focus only on players that changed teams between two years; that is, they played for exactly one team in year x and exactly one different team in year x+1. This is of course a strong simplification, but it works because moving during free agency is essentially the only way a player can play for only one team in one year and another the next, so

we can assume that the two lists of players are mostly equivalent. Importantly, because the analysis is focused exclusively on players joining other players, irregardless of location, the choice to omit the players who stayed with their team in free agency does not affect our analysis. From here, it was just a matter of writing another Python script to generate data points and then modify and consolidate the column data from a variety of sources into a single input matrix format with a total of 82365 entries.

IV. Definitions

For the sake of clarity in the remainder of this paper, I will address some possible ambiguities. First, there is no reasonable way to accurately document the relationship of every pair of players in the last two decades. Thus, I narrowed down the focus of this project to finding quantifiable and accessible features that could possibly affect the players' choices. In the remainder of this paper, the "closeness" of two players will be directly related to the likelihood that they would end up on the same team the following year if given the opportunity, as opposed to how close of friends they are in real life.

In addition, the input data includes 3-pointer volume and free three volume "incompatibilities", which I define using the following formulas, respectively:

$$t = (t_1 - 1.5)\,(t_2 - 1.5) \qquad\qquad f = (f_1 - 4)\,(f_2 - 4),$$

where t_i and f_i are the three-point attempts and free throw attempts per 36 minutes for each player i. The idea is that certain players are considered sharpshooters or ball-dominant players and thus have a high volume of three-pointers and free throws, respectively. However, teams usually tend not to have multiple sharpshooters or multiple ball-dominant players, because their play styles often conflict and result in inefficient offense. Thus, higher values of three-point incompatibility result when two players either both shoot threes at a high volume or both do most of their work away from the perimeter, which can clog up the paint for the other players. The reasoning for free throw incompatibility follows similarly.

Finally, "positive outcomes", "successful data points", and "1 labels" will be used interchangeably to refer to two players entering free agency and ending up on the same team the following year.

V. Results — Regression Analysis

The first task was to find which input columns were the most relevant to evaluating whether a pair would join the same team. For this, the greatest challenge was the imbalance of the data labels: roughly 97.5% of the labels were 0. Thus, running logistic regression as is would result in predictions of all 0's, and the resulting 97.5% accuracy wouldn't be helpful at all. One common way to mitigate this issue is to construct artificial '1'-labeled points based on the existing data, but because this dataset is so noisy, I believe that this approach would do more harm than good. I eventually decided on a "mini-batch" approach of sorts: I generated many temporary "sub-datasets" containing all of the 1-labeled data points and an equal number of randomly sampled 0-labeled data points, and then ran logistic regression on these sub-datasets. For each iteration, I extracted and stored the weights, and at the end I took the mean of the stored values as my "average regression" coefficients. From there, I applied the weights and sigmoid function to the entire dataset to generate my final predictions. This "average" approach was conceived to take advantage of the Law of Large Numbers and assumed that as the number of sub-datasets increased, the averaged coefficients would converge to the "true" values.

As a sanity check, let's first examine the coefficients that our approach generated:

Variable	Coefficient
(Intercept)	0.01145038
years as NBA teammate	-0.01789091
(years as NBA teammate)^2	0.02202203
age difference	0.00823173
(age difference)^2	-0.00749916
same last name	0.17533091
same jersey number	-0.00717772
same college alma mater	-0.04425542
same international region (USA vs. nonUSA)	0.06647378
3-pointer volume incompatibility	-0.00733215
free throw volume incompatibility	-0.00343217

Figure 1: "averaged regression" coefficients

Fortunately, all of the coefficients appear reasonable. For instance, the coefficients corresponding to 3-pointer and free throw incompatibility have negative signs, which matches our hypothesis for why we included the two variables in the first place. In addition, the plots of the contributions of both teammate years and age difference aligns with expectations: being teammates for many years facilitates a strong

relationship, and being around the same age allows players to mesh better on the court and bond more off the court.

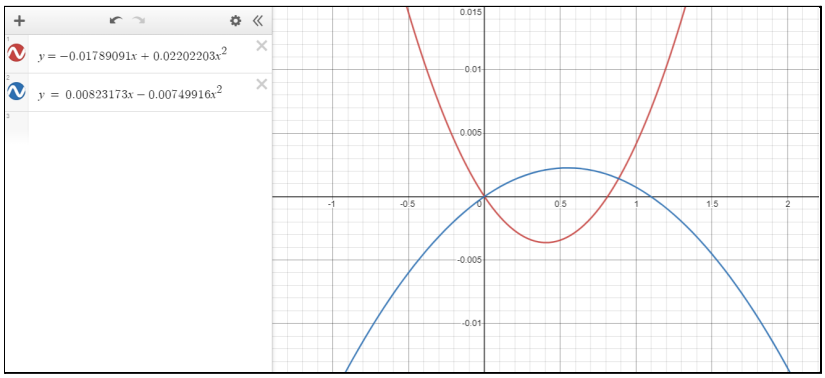

Figure 2: visualization of quadratic coefficients

To me, the most surprising result was the dummy regarding having the same last name, with a relatively large coefficient of 0.17. However, upon further inspection, there were only 164 data points where this condition was true, which implies that overfitting was occurring to some degree.

This evokes the question of which of these coefficients can be considered significant. When I ran the regression iterations in R, I could see which inputs were considered to have a real effect for that particular instance. Over many iterations, I noticed that the 3-pointer incompatibility coefficient was regularly denoted as statistically significant; the coefficients for teammate years and age difference sometimes were as well.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.027498   0.039873   0.690   0.4904
NBA_Years     -0.069024   0.115848  -0.596   0.5513
NBA_Years_2    0.036187   0.031970   1.132   0.2577
AgeDiff       -0.002261   0.043628  -0.052   0.9587
AgeDiff_2     -0.005494   0.010266  -0.535   0.5925
Last_Name      0.316625   0.541877   0.584   0.5590
Jersey         0.120825   0.204275   0.591   0.5542
College       -0.231781   0.242385  -0.956   0.3389
ThreePointRate -0.008098   0.003914  -2.069   0.0386 *
FreeThrowRate  0.003625   0.008498   0.427   0.6697
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: statistical significance of coefficients (R output)

Returning to the model, the results were optimistic. Out of the 82000+ data points, the averaged model predicted a positive result 53.54% of the time and a negative result 46.46% of the time, which is close to the 50-50 split expected from the approach I used. However, due to the natural distribution of the dataset labels, the fact that I did not predict '0' 97.5% of the time removes the possibility of achieving a good raw accuracy (by any possible definition of "good"). The best I can hope for is that the model was able to perform better than random chance; that is, if I were to blindly predict '1' with probability 53.54%, what are the chances of achieving the same or better results as the trained model?

Because the random chance model predicts blindly, we can compare the results for the 1-labeled and 0-labeled data points separately. In total, there were 2048 positive data points, and the model correctly predicted 1133 of them, giving an accuracy of 55.32%. On the other hand, there were 80317 negative data points, with the model correctly predicting 37354 of them and yielding a 46.51% accuracy. These accuracy values aren't very reassuring, but the results were still notable: for the positive data points, the random model would have performed at least this well with probability of only 5.50%. While this isn't technically statistically significant, it provides optimism that these specific variables indeed provide some insight into players' free agency decisions.

Thus, for the remainder of this paper, I will calculate the "closeness" score as the pre-sigmoid linear combination value using these coefficients, normalized to be between 0 and 1; that is:

$$s_i = \frac{(W \cdot x_i + b) - s_{min}}{s_{max} - s_{min}}$$

where s_{min} and s_{max} are the minimum and maximum calculated values, respectively. After normalization, the average score is about 0.4152, with a relatively narrow standard deviation of 0.033.

VI. Results — Triadic Closure

Now that I have weights, I can explore the main focus of the project: whether triadic closure has an effect on our output labels. First, I wanted to confirm that our calculated scores translated well to a model evaluating trios instead of pairs. To do so, I needed to determine a good subset of players to test this on. Eventually, I decided on triples of players who were all entering free agency at the same time, played on different teams the year before, and eventually changed teams during that free agency period. This subset was desirable because it generated triples of players in which all three players were on the move, which

reduced the effects of team-specific factors on our data. In total, there were more than 84000 trios of players from 2006-2016 that matched that description.

To perform our check, I iterated through all trios where exactly one of the three pairs in those trios was a positive outcome and then noted the relative scores of those three pairs. Fortunately, the results somewhat matched expectations: the positive pair was the highest scoring 34.8% of the time, compared to 32.6% and 32.7% for the middle and lowest scoring pairs. This indicates that our "averaged regression" model coefficients hold for both pairs and trios.

In addition, the highest scoring pair theoretically is more likely to end up on the same team than random chance; that is, better than a 1 in 30 chance. Interestingly, about 3.57% of the trios were successful, which is significantly better than the 3.33% that was expected. In a randomized trial with 1/30 probability, this outcome of 2529 successful out of 70851 would only occur 0.03% of the time, so this result was quite encouraging. Bucketing across years, the results are very volatile with percentages ranging from 1.9% to 6.2%, so it's unlikely to find any worthwhile insight there.

Year	Successes	Total	Percentage
2006	27	1402	0.0193
2007	86	3404	0.0253
2008	140	2478	0.0565
2009	108	2681	0.0403
2010	120	5994	0.0200
2011	20	1016	0.0197
2012	615	17546	0.0351
2013	473	11466	0.0413
2014	454	7366	0.0616
2015	342	8444	0.0405
2016	144	9054	0.0159
Overall	2529	70851	0.0357

Figure 4: prediction success across the years

With the confirmation out of the way, I now explored the network effects aspect. In particular, I focused on the two higher scores of each trio. If triadic closure were to have an effect on whether players ended up on the same team, we would note a significant difference in those scores between positive and negative data points. To compare, I generated a pivot table to compare the geometric mean of the two higher scores across the two possible labels:

Label	Count	Average Score	Score Deviation
0	68469	0.424011	0.024921
1	2382	0.424110	0.024342
Grand Total	**70851**	**0.424014**	**0.024902**

Figure 5: scores of other two pairs for pairs that joined vs pairs that didn't

At first glance, the results appear promising, as the average score is higher for the '1' label. However, two-sample hypothesis testing yields a z-score of just 0.19, leading to an unspectacular p-value of 0.423.

Another way with which I approached triadic closure was to filter out trios where the two higher scores were both above various thresholds and to see whether the frequency of positive outcomes for the least close pairs is significantly higher than the rest of the population. For a few thresholds between 0.42 and 0.45, I compared the raw success of trios that met or didn't meet the threshold, keeping in mind the benchmark random probability of 1 in 30, or 0.333%. Theoretically, the raw success rates would be higher for trios that met the benchmark requirements.

Threshold	Proportion of successes		p-value
	0	1	
0.42	0.0337	0.0335	0.484
0.425	0.0343	0.0326	0.349
0.43	0.0341	0.0324	0.355
0.435	0.0339	0.0326	0.412
0.44	0.0339	0.0313	0.344
0.445	0.0341	0.0281	0.225
0.45	0.0339	0.0286	0.288

Figure 6: raw success rate for various thresholds

Unfortunately, nothing here can be considered statistically significant. Notably, however, the success rate counterintuitively is lower for the trios that meet the threshold across the board, although the difference is negligible. There also appears to be a steep drop-off between 0.44 and 0.445, but this can be attributed to the sample size decreasing by about 50% between those two checkpoints.

To wrap up this analysis, I attempted one last test: I filtered out trios in which exactly one of the pairs had a positive outcome to see how often the lowest pair was the one that was successful. In theory, the regression results from earlier imply that this value should be lower than 1/3, as the other two pairs have higher scores and should be more likely to have positive outcomes. However, triadic closure could mean

more successes for the pairs with the lowest score within each trio at the expense of the higher scoring pairs, and the results were actually incredibly interesting:

Threshold	Proportion of successes		p-value
	0	1	
0.42	0.3269	0.3266	0.511
0.425	0.3237	0.3313	0.254
0.43	0.3190	0.3472	0.013
0.435	0.3200	0.3597	0.005
0.44	0.3246	0.3437	0.146
0.445	0.3285	0.3017	0.890
0.45	0.3291	0.2789	0.977

Figure 7: relative success rate for various thresholds

There are two major takeaways from this chart: one, that, for a sweet spot around 0.435, there's a statistically significant difference for the trios that had two pairs of players above the threshold; and two, that there's also a statistically significant difference for the trios that had two pairs of players score above 0.45, with abysmal numbers mirroring the ones in Figure 6. The first result is exactly what one would expect if triadic closure is a thing, but the second result casts some doubt as to why the proportion of successes has a quasi-quadratic relationship with the threshold values.

I thought a lot about possible theories why, but I couldn't figure out any valid reasons how a clear trend emerges in such a noisy dataset, especially when earlier tests appear to indicate no effect. One possibility I considered was that some of the inputs to the linear regression might have been adversarial in trios: that is, for example, players that are both highly 3-pt-incompatible with a third player are more likely to be compatible. However, the numbers don't back this, as the average lowest score fluctuates minimally between different threshold values (standard deviation is about 0.02).

Threshold Range	Average Lowest Score
0.42 - 0.425	0.403
0.425 - 0.43	0.401
0.43 - 0.435	0.397
0.435 - 0.44	0.397
0.44 - 0.445	0.404
0.445 - 0.45	0.411
> 0.45	0.402

Figure 8: average lowest score for various threshold ranges

I'm not sure how much valuable insight I can expect to gain from applying a noisy regression model to a noisy dataset of trios of players, but I think it may be worthwhile in the future to reexamine this trend and perhaps show mathematically whether it's trivial or not.

VII. Limitations

There were a few noteworthy shortcomings in this analysis. First, the dataset was extremely sparse and noisy, so it was very difficult to extract actionable insights on player decisions, as expected. However, there were still some meaningful findings that could be examined more in detail in future analysis. In addition, the vast majority of the analysis did not take longitudinal trends into account, such as the rise of 3-point shooting in the past decade or so. It could be worthwhile to segment the data by year to more accurately describe the state of the league. Finally, in the data acquisition and modification phase, the inputs to the logistic regression had about 11000 unusable data points, about 10% of the total scraped dataset. However, the missing data did not appear to be related, so it likely did not have any effect on the analysis.

VIII. Acknowledgments

The data for this project was acquired from Basketball Reference, as well as pre-cleaned Kaggle datasets thanks to Wyatt Welsh and Fernando Blanco. Links are below:

- https://www.basketball-reference.com
- https://www.kaggle.com/wyattowalsh/basketball
- https://www.kaggle.com/whitefero/nba-players-advanced-season-stats-19782016

Also, I'd like to take this opportunity to thank Miki and the rest of the course staff for a great semester! ORF387 was a wonderful class, and I really learned a lot. I definitely plan on taking more courses with Miki, and I hope we can meet again in the future!