Mackenzye Leroy
William McDevitt
Edwin Purcell

Bayesian Machine Learning Project Report

1. Introduction

For our Bayesian analysis, we used Bayesian regression to attempt to predict whether a shot was made based on distance from the hoop using basketball data. We started by getting our data from Kaggle of the 2016-2017 season-specifically we focused on the Cleveland Cavaliers and the Houston Rockets. The Cavaliers at the time were a very solid all-around team while the Rockets were more known for shooting which we thought would make an interesting comparison when forming our predictions. The dataset itself contains information about the home team, the away team, and the x- and y- coordinates of the player on the basketball court when attempting the shot with (0,0) corresponding to the lower left-hand corner of the court. It also contains a wide range of other predictors like time the shot was taken, the result of the last shot, and the type of shot attempted (e.g dunk versus jumpshot), but we chose to simply focus on distance as it was the singlest biggest predictor in every other attempt we saw to predict shot outcome.

Before creating our models we cleaned our data by filtering out the players that shot less than 25 times over the course of the season since players who shot fewer than 25 shots would be hard to predict given their small sample size. We then used the x and y locations of our players to calculate their Euclidean distance from each basket to get an estimate of how far away the shot was from the rim. We thus had two distances with each one corresponding to one of the baskets and assumed that the basket closer to the player was the one he was shooting at since the vast majority of shots would be towards the closest basket. Next we quickly combined blocked shots with missed shots because we only care if the shot was made or not. After cleaning, we had 6926 and 7017 shot attempts with 47.1% and 46.2% of the attempts made for Cleveland and Houston respectively. For each team, we held out 1000 shots for testing and built our model on the remaining shots (5926 for Cleveland and 6017 for Houston). While in the past, most models struggled to predict the shot accuracy of players who do not attempt many shots, we believe that utilizing hierarchical models will give us better predictions for these types of players and an overall better accuracy.

2. Mathematical Linkage between Problem and Methodology

To predict whether a shot is made or missed, we need to calculate the probability of each outcome happening. To do this, we can utilize Bayesian logistic regression, which is a commonly used approach to calculate probabilities of categorical response variables. Additionally, we can

potentially increase the overall accuracy of our model by separating the predictions based on who is taking the shot. Different players in the NBA specialize in different things and someone like James Harden, a player known for taking a high volume of 3pt jump shots, will make more deep shots than a player like Clint Capela who spends all of his time around the basket. Therefore, a Bayesian approach such as pooled and unpooled models could be useful in observing some of the differences between the trends we expect to see from different players.

3. Bayesian Methods

For each model, we used both sampling and variational approximation approaches to find the posterior distribution so we could compare their results. The first model we ran was a pooled model for each team (Figure 3) which looks at overall how likely any player would be to make a shot based on distance. As can be seen, we utilized logistic regression to model our output. Here, we used gaussian priors on our predictors, a deterministic equation to where our only predictor variable was distance, and a Bernoulli distribution to calculate the output where a 1 represents a made shot and a 0 represents a missed shot. We utilized non-informative gaussian priors for our predictors since we assumed the amount of data we had would cause our posterior mean to converge to the sample mean.

We started off using a pooled model, which uses all the observations from every player to train a single model to make predictions. Next, we used an unpooled model to break down the predictions based on individual players in each team. Unpooled models use unique betas for each player. We chose this type of model with the hope that conditioning on the player shooting would increase our chance of correctly predicting the shot outcome (Figure 4). Again we used gaussian priors and realize that this may have an impact on the players that shot a relatively low amount of shots. Finally, to attempt to correct some of our outliers with a positive slope, like Bobby Brown and Jordan McRae, we used a hierarchical model. Hierarchical models work by incorporating all observations from each player into the betas for each individual player with the hopes of shrinking some of the extreme estimates we can get for players that take a low number of shot attempts towards their respective team averages.

4. Results

Pooled Model

After running our pooled model, we were not surprised to see a rather steady decreasing slope meaning that as distance from basket increased, the chance of a shot being made goes down linearly with essentially no difference between sampling and variational approximation and minimal difference between the two teams as seen in Figures 1 and 2. Results for our models can

be seen in Figure 12. Our pooled model with HMC had an accuracy of 58.8% for the Cavaliers and 62% for the Rockets and with. These accuracy scores are relatively good considering we would have an accuracy of 53% for the Cavaliers and 54% for the Rockets if we just guessed miss for every single shot, and we only used one predictor.

Unpooled Model

Figures 5 and 6 show the resulting regression lines for each player on each team using sampling and an unpooled model. As we can see in these figures, almost every player on both teams has a negative slope with one player on each team with a positive slope. The positive slopes resulted from players who had very few shots throughout the season, but made quite a few three pointers in those limited shots. The players that have the steepest negative slopes are the respective team's centers which makes sense given that the majority of the shots that they make are very close to the basket. Note that Figures 5 and 6 are the results of sampling, but much like in the unpooled model, we saw almost no distinguishable difference between the plots for sampling and variational approximation. As far as predictive abilities go, we get a slight increase in accuracy when we use the unpooled models with 61.6% and 62.1% for the Cavaliers and Rockets respectively using sampling. Again these are very solid improvements from our 53-54% baseline with only one predictor.

Hierarchical Model

Figures 8 and 9 show the resulting regression lines for each player with our hierarchical model using a sampling approach. As expected, we see that the sampling method resulted in the trendlines for all players shrinking to the pooled model trend especially for players with a low number of shots taken. One of the most interesting results of our hierarchical model was that unlike with both the pooled and unpooled model, we see a pretty drastic difference in resulting regression plots depending on our method. Figures 10 and 11 show the resulting regression plots of our hierarchical model using variational approximation. Both figures resemble our unpooled models more than they resemble the hierarchical model we got from sampling, indicating that variational approximation did not shrink our small sample outliers the way we expected. We credit this to variational approximation being less accurate than sampling for complicated distributions. Again, we get another slight increase in predictive ability with our hierarchical models with an accuracy of 61.7% for the Cavaliers and 62.9% for the Rockets using sampling. As potentially expected given our plots though, we don't see much of an improvement from the unpooled model to the hierarchical when we look at the variational approximation models.

Overall Trends and Notes on Results

Figure 12 summarizes the predictive results of our models. Overall we tend to see better predictions as we move from a pooled model to an unpooled model to a hierarchical model. Sampling also tends to outperform variational approximation, especially with our hierarchical models. Of course, we could see slight variations in results if we ran our models again due to the stochastic nature of the sampling and variational methods used.

5. Conclusions

For each respective team, the hierarchical models performed the best (when using accuracy metric), with less than a 3% increase in accuracy for both teams when compared to the pooled model. We can conclude that distance from the basket is a decent predictor of the shot outcome. Since hierarchical models performed well, it was worthwhile to separate shots out by who is shooting the ball. This conclusion makes sense when considering players on the same team have different skill sets namely three point shooters, mid range shooters, and centers.

Shortcomings and Future Work

Future directions could be to add other predictors such as previous distance from defender, shot made/missed, time left in the game, etc. We also could adjust our priors to be a multimodal distribution where each mode corresponds to a sweet spot on the court where players are likely to make their shots.

Appendix

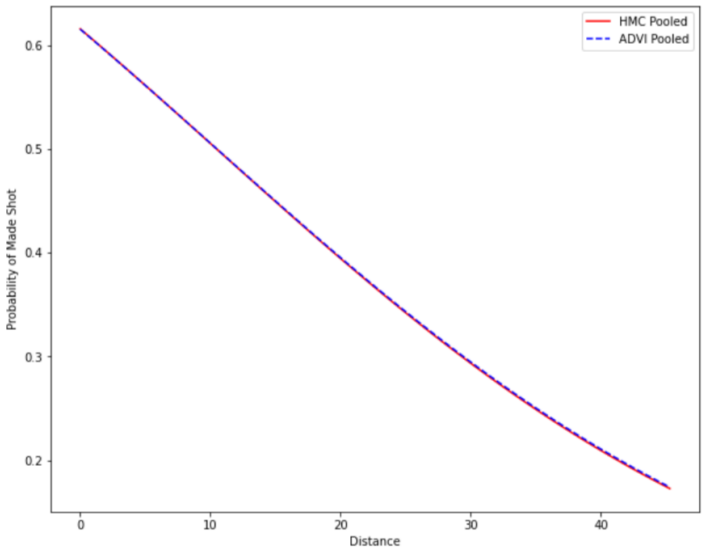

Figure 1: Pooled Model Result for the Cleveland Cavaliers

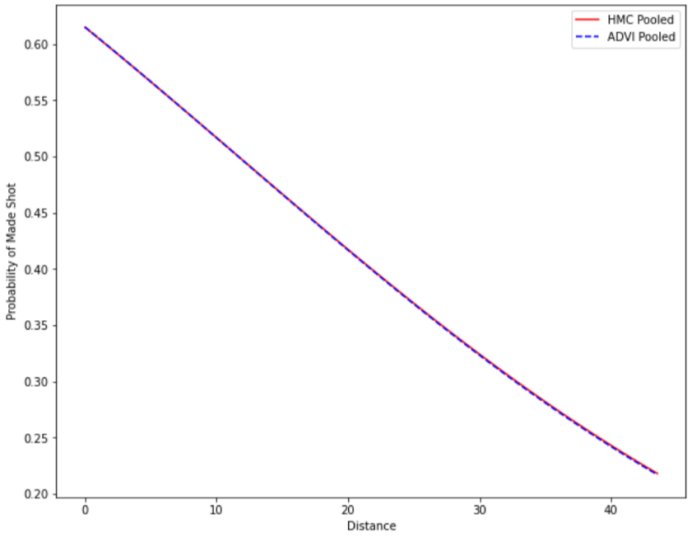

Figure 2: Pooled Model result for the Houston Rockets

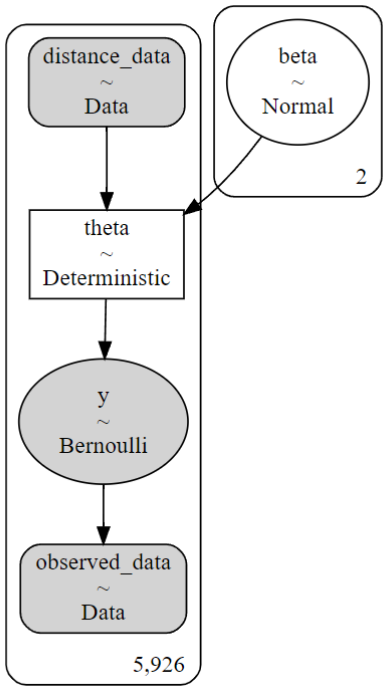

Figure 3: Pooled Model Diagram for the Cleveland cavaliers. Note that the diagram for the Houston Rockets is identical apart from the number listed on the plate (6,017 instead of 5,926)

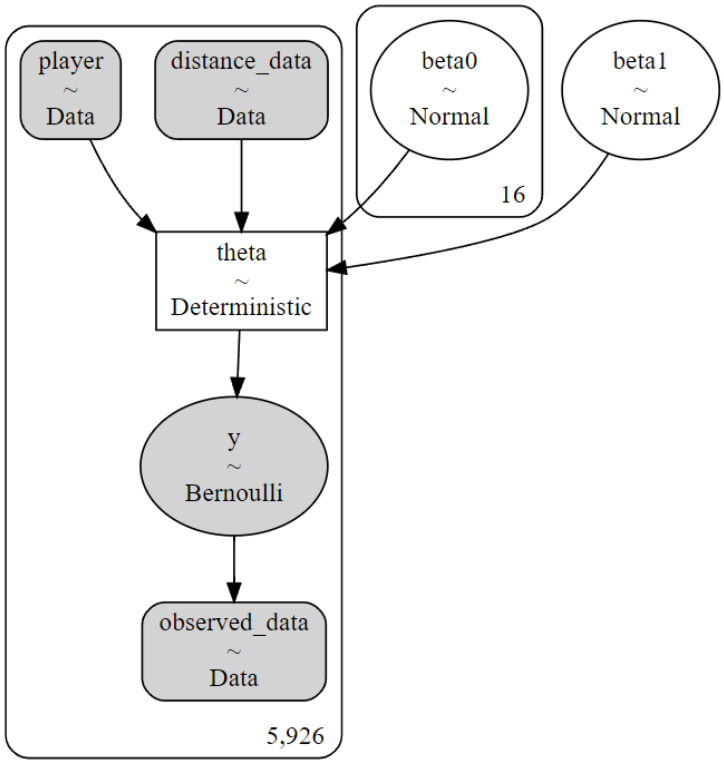

Figure 4: Unpooled Model Diagram for the Cleveland cavaliers. Again the only difference for the Houston Rockets is the value on the plate.

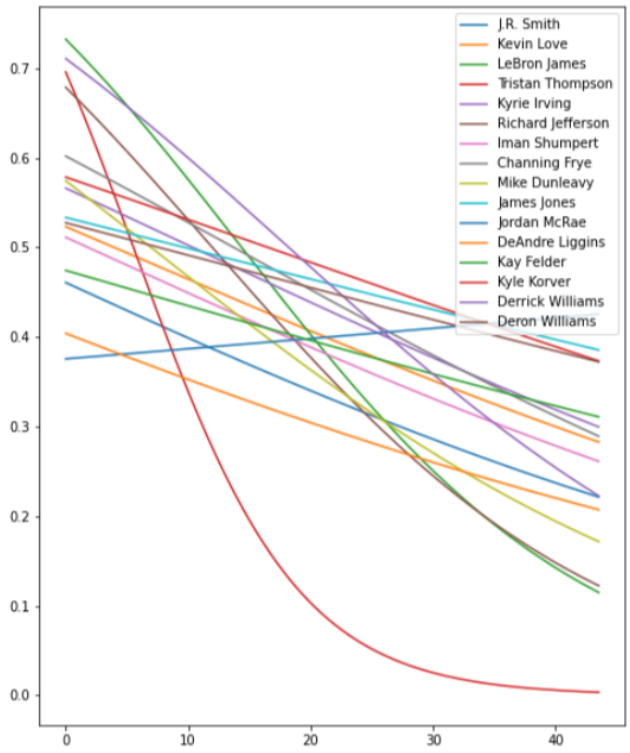

Figure 5: Unpooled Model Result for the Cleveland Cavaliers

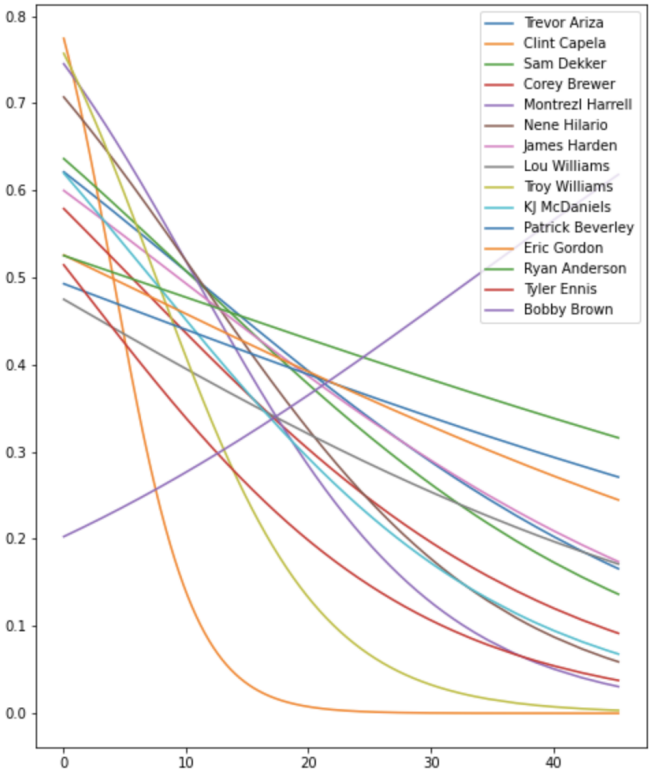

Figure 6: Unpooled Model result for the Houston Rockets

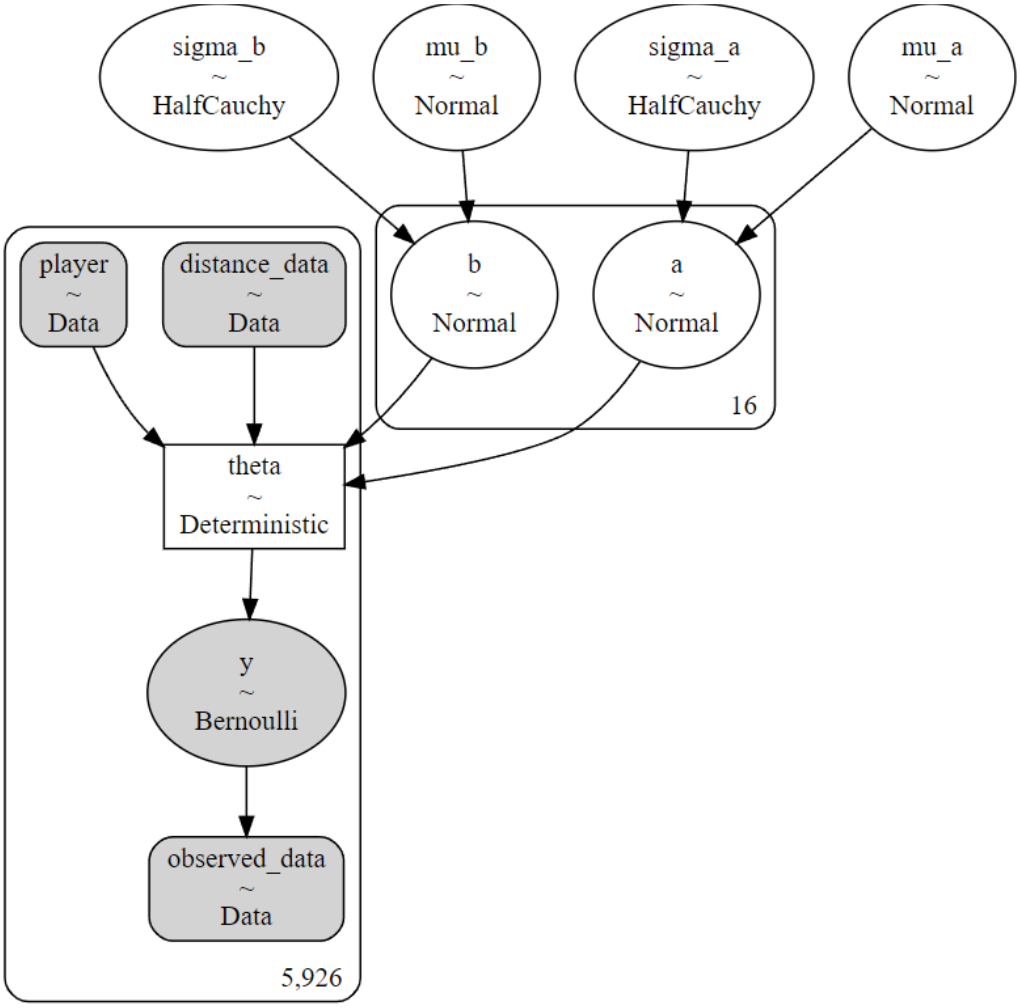

Figure 7: Hierarchical Model Diagram for the Cleveland Cavaliers. Like the pooled and unpooled model, the only difference for the Houston Rockets is the value on the plate.

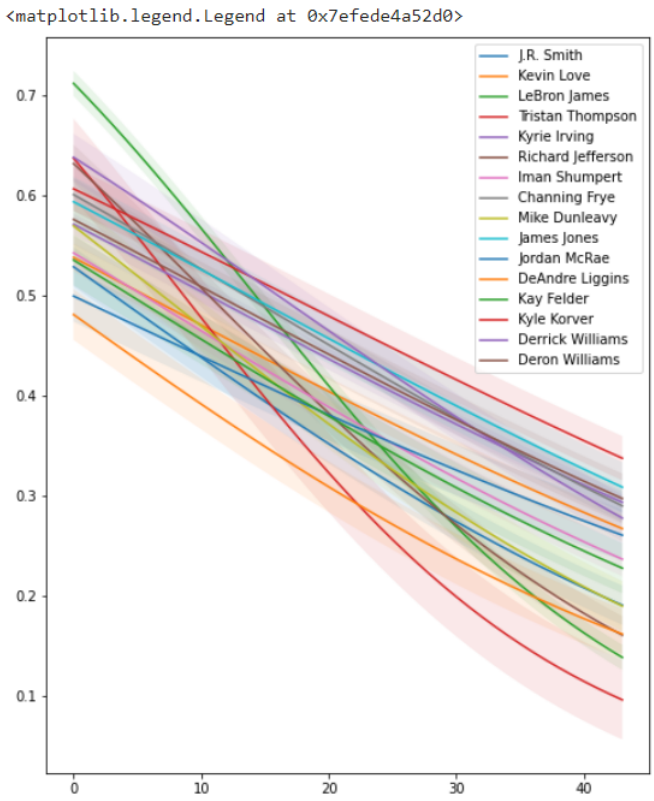

Figure 8: Hierarchical Model Result Using Sampling for the Cleveland Cavaliers

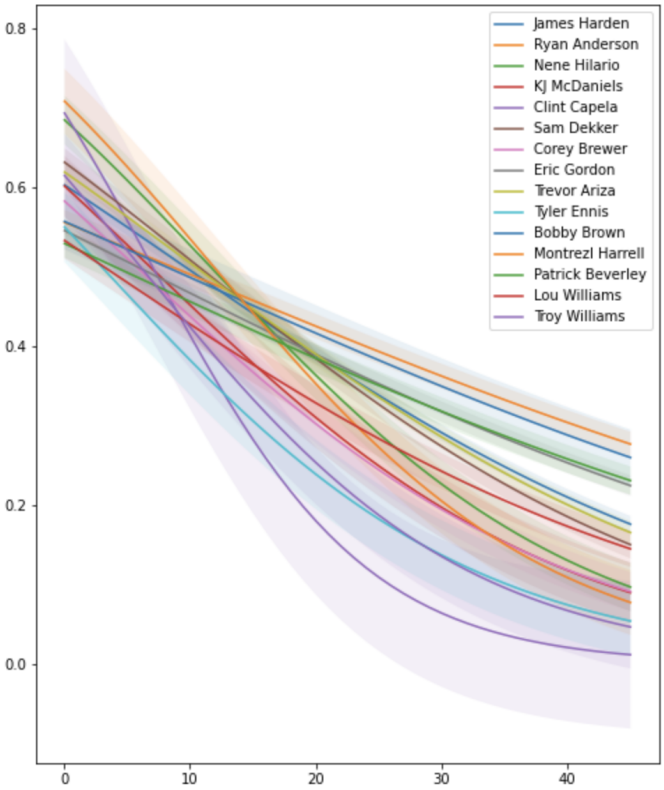

Figure 9: Hierarchical Model Result Using Sampling for the Houston Rockets

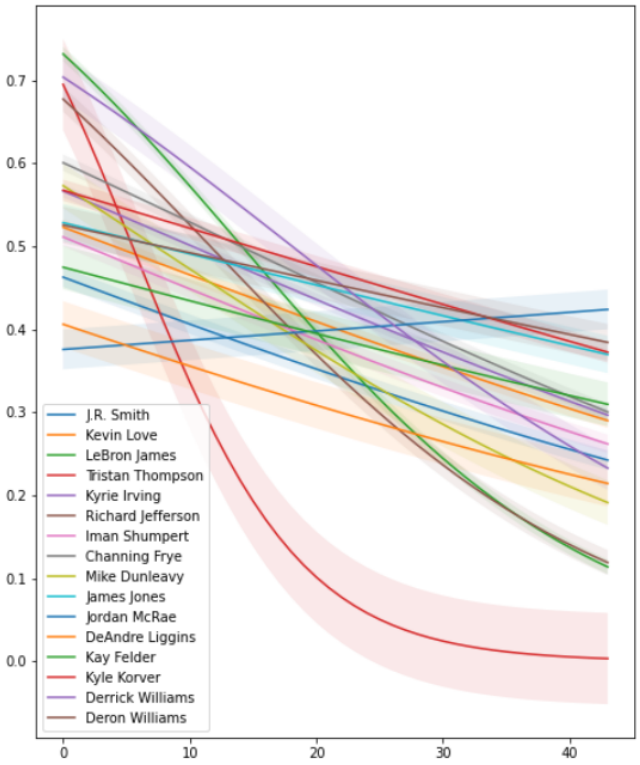

Figure 10: Hierarchical Model Result Using Variational Approximation for the Cleveland Cavaliers. Note the lack of shrinkage.

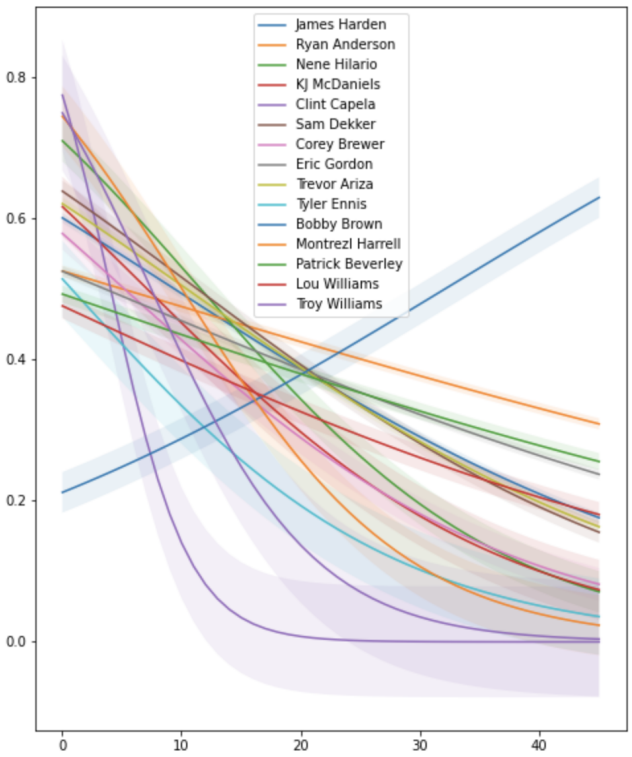

Figure 10: Hierarchical Model Result Using Variational Approximation for the Houston Rockets. Note the lack of shrinkage

Team	Model	Method	Accuracy	Recall	Precision
Cleveland Cavaliers	Pooled	HMC	58.80%	49.90%	55.20%
		ADVI	58.30%	50.10%	54.50%
	Unpooled	HMC	61.60%	38.50%	62.70%
		ADVI	61.50%	38.00%	62.70%
	Hierarchical	HMC	61.70%	51.70%	61.70%
		ADVI	60.60%	46.00%	61.50%
Houston Rockets	Pooled	HMC	62.00%	56.30%	60.90%
		ADVI	62.00%	56.50%	60.90%
	Unpooled	HMC	62.10%	48.50%	63.30%
		ADVI	62.90%	47.90%	65.00%
	Hierarchical	HMC	62.90%	52.70%	63.20%
		ADVI	62.20%	47.70%	63.80%

Figure 12: Final prediction results of every model. Note that due to the stochastic nature of both sampling and variational approximation these results may vary slightly on different runs, and would very likely change with a different training/test split.