097400 – Casual Inference Project – "Hot Hand"

Ariel Abramowitz 205552599
Gil Weissman 322593377

August 2022

Link to Repository

Introduction of the problem

The "hot hand" is a phenomenon, previously considered a cognitive social bias, that a person who experiences a successful outcome has a greater chance of success in further attempts. The concept is often applied to sports and skill-based tasks in general and originates from basketball, where a shooter is more likely to score if their previous attempts were successful.

Gilovich, Vallone, and Tversky[1] , in which the authors found that that the “hot hand” momentum does not exist, despite the belief among basketball fans and experts that there is momentum in shooting performance. In the years since the seminal paper was published, a consensus has emerged that the hot hand is a “myth”[2][3]. Some recent studies[4] have observed evidence for the "hot hand" in some sporting activities, and found that some previous didn't observed it due to biased selection. However, other recent studies[5][6] have not observed evidence of the "hot hand".

In our project we examine whether the "hot hand" effect exists or not in the NBA.

Our causal question is: **Does scoring/missing the previous shot towards the basket during a basketball game affect the probability of a future score on the next shot by the same player later in the game?**

A player's shot result will be used as a treatment (T) for his next shot (in the same game) and as an outcome (Y) for his previous shot. That's why we are interested in the shooting result starting from the player's second shot. This means that for each player and in each game, we would like to examine the causal question we described, starting from the second shot in the game to the last shot. For this purpose, we arranged the data by game and by player, which means that for each game we have all the shots taken in the game ordered by chronological order for each player.

We denote shots which their previous shot (by the same player) was successful as the treatment group, and the other shots as the control group.

We measure the "hot hand" effect using the ATE index as we learned in the course. Note that some previous studies measured other indexes such as ATT.

We consider $ATE = 0$ as the null hypothesis, because of the old consensus on this question.

Data

The data we use to answer the question is a dataset which was created by two students of the course (Alon Tsaizel and Liran Halperin) a year ago. We used this dataset after getting their and Rom's approval. This dataset consists of more than 100,000 shots taken by 220 different players, during 2014-2015 NBA season. The dataset was created by applying join actions between 3 datasets:

1. Dataset from Kaggle website with data that includes shots by players and potential confounders. Below is the description of the dataset:

   Data on shots taken during the 2014-2015 season, who took the shot, where on the floor was the shot taken from, who was the nearest defender, how far away was the nearest defender, time on the shot clock, and much more.

   This dataset is the main one (out of the three) and includes most of the features we needed for our project.

2. Dataset from the basketball reference website that includes physical features of each of the 220 players (e.g., player's height and weight).
3. Dataset from the basketball reference website that includes statistical measures of each of the 220 players' performances from the previous season (2013-2014).

The dataset consists of 102692 shots, taken in 1808 games, by 219 players, so 11290 game-player combinations are recorded.

Following an initial investigation of the problem, we decided to include the following 18 confounders.

Confounders from the first dataset:

- LOCATION - indicator whose value is 1 when the game is a home game and 0 otherwise.
- SHOT_NUMBER - the shot number of the throwing player in the game.
- PERIOD - the quarter in which the shot was thrown.
- GAME_CLOCK - the number of seconds left on the quarter clock at the time of the shot.
- SHOT_CLOCK - the number of seconds left on the shot clock at the time of the shot.
- DRIBBLES - the number of dribbles by the throwing player before the shot.
- TOUCH_TIME - the number of seconds the throwing player held the ball before the shot.
- PTH_TYPE – the type of the shot.
- SHOT_DIST - the distance of the shot from the basket.
- back2back - indicator whose value is 1 when the thrower played in a game the day before the current game and 0 otherwise.
- CLOSE_DEF_DIST – distance of the thrower from the nearest shield.

Confounders from the second dataset:

- player_weight – weight of the player (Kg)
- player_height – height of the player (Sm)
- defender_weight – weight of the defender (Kg)
- weight_height - height of the defender (Sm)

Confounders from the third dataset:

- prev_3P% - Last season 3-point shooting percentage.
- prev_FT% - Last season free-throw shooting percentage.
- prev_FG% - Last season field-goal shooting percentage.

All of these covariates are included in the dataset. Therefore, we can consider all the features described above (X) as confounders except for the treatment (T) and the outcome (Y). We assume that there are no hidden confounders that we have not specified.

After removing the first shot of each player in each game and all NAN values, we were left with 59552 records. 45.25% of the records were from the treatment group.

We normalized X using min-max normalization. That scaling may improve the performance of some of the models we used, leading to a more accurate ATE estimation. Not scaling the data in distance-based algorithms (e.g. KNN) results in uneven weighting for each feature.

Not using PTS column

The PTS column represents the number of points added to the team as a result of the shot that was taken (0, 1, 2 or 3)

For the classification task of predicting the binary class of the shot result (in / out) the PTS column includes the answer. A simple decision rule for this classification task, including the PTS feature, would be 0 if PTS=0 and PTS=1 elsewhere. This basic decision rule would result in 100% accuracy. Therefore, we decided not to use this feature in our model.

Looking at the casual graph, Y and SHOT_DIST are the only parents of PTS in the graph, and therefore cannot be used as a confounder in our model.

Data Exploration

Figure 1 displays the distribution of propensity scores for the treatment and control groups.

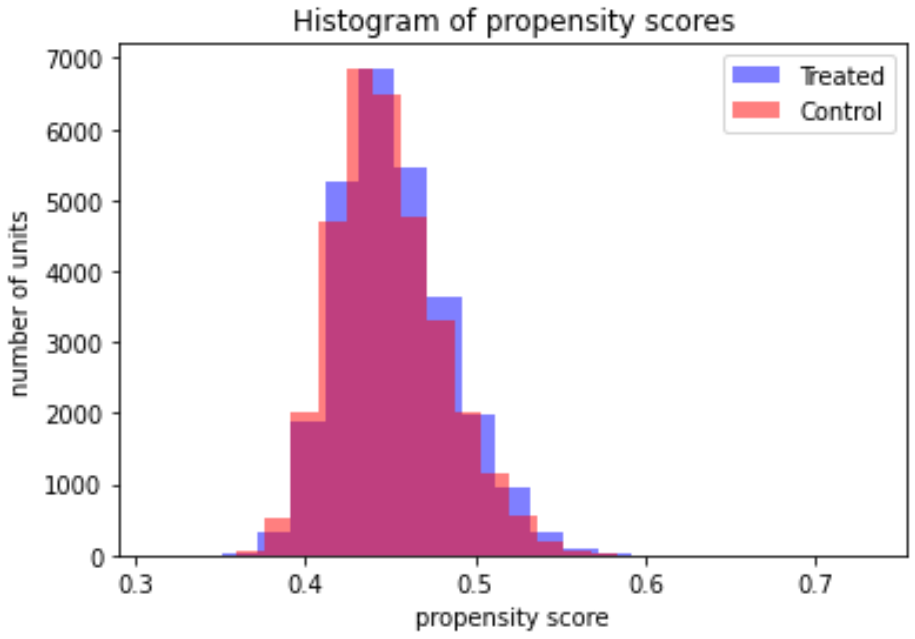

Figure 1 – Histogram of propensity scores

It can be seen from *figure 1* that there is an overlap between the groups almost all over the range of the propensity scores.

In addition, we measured the correlations between the different variates as can be seen in the following *figure 2*.

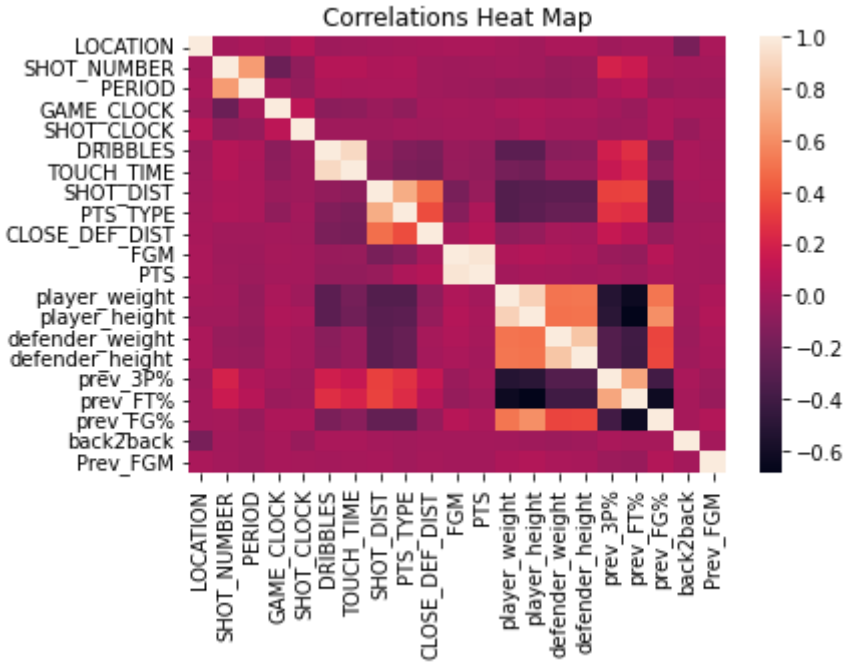

Figure 2 – Correlations between variates

No meaningful conclusions have been found in *figure 2*, therefore, we decided to keep all the variates (except PTS).

Assumptions

SUTVA

It is very reasonable to assume that shots taken at some game does not affect potential outcomes of shots taken at other games. That is because people tend to be influenced mostly by current events related to them, and that games take place at long time differences from shots of other games. Thus, it is likely that events that affect the game occur within the same game.

It is also likely to assume, although with less certainty, that shots by different players do not affect each other. Furthermore, while a player's shot may affect the result of his subsequent shot, it does not directly affect the shots afterwards. This is again, because people tend to be influenced mostly by the most current event related to them. That is, we assume that given the previous shot of a player and given the confounders X, the potential outcome of the next shot does not depend on any other shot (certainly not on future shots). Therefore, we can assume SUTVA with some degree of confidence.

Consistency

It is reasonable to assume that $Y = TY_1 + (1 - T)Y_0$, i.e., the shot's outcome equals the shot's potential outcome had the player's last shot scored, if the player's last shot actually scored, and equals to the shot's potential outcome had the player's last shot missed, if the player's last shot actually missed.

No unmeasured confounders

We assume that given the previous shot's result, only the confounder mentioned earlier matter for predicting the next shot's result. This is a fairly reasonable assumption because the dataset includes statistics concerning the moment of the shot, and these features are commonly measured in basketball, even though there may be a (negligible) effect of unmeasured features.

Common support

As shown above in *figure 1*, the estimated propensity score of every sample is greater than 0 and smaller than 1, thus, assuming that our propensity estimator is correct, there is a common support.

Methods

We used the following methods for estimating ATE:

- IPW – In this method each unit is weighted using the inverse of its propensity score estimation. The propensity score of unit i is defined as $P(t_i = 1|x_i)$ which is the probability of the unit to get the treatment (score in the previous shot) given all its confounders.
  The estimation of ATE is
  $$\widehat{ATE} = \left(\sum_{i=1}^{n} \frac{t_i}{\hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{y_i t_i}{\hat{e}_i} - \left(\sum_{i=1}^{n} \frac{(1-t_i)}{1-\hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{y_i(1-t_i)}{1-\hat{e}_i}$$
  where $\hat{e}_i = \hat{P}(t_i = 1|x_i)$ is the propensity score, estimated using logistic regression.

- S-Learner using gradient boosting, i.e., a model $f(X,T)$ for estimating Y. The estimation of ATE is
  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)$$
  $f(X,T)$ was trained with X and T as the training set and Y as the test set.

- T- Learner – using gradient boosting, i.e., two estimators, $f_1(X)$ for the treatment group and $f_0(X)$ for the control group. The estimation of ATE is
  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} f_1(x_i) - f_0(x_i)$$
  $f_1(X)$ was trained with X of the treatment group as the training set and Y of the treatment group as the test set.
  $f_0(X)$ was trained with X of the control group as the training set and Y of the control group as the test set.

- Matching – using 1-Nearest-Neighbor model for matching. In this method for each unit, we estimate the potential outcome of the "opposite" treatment using the most similar unit from the "opposite" group. For example, for a unit from the treatment group we will find its nearest neighbor from the control group.
  The estimation of ATE is
  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} ITE(i)$$
  where
  $$ITE(i) = \begin{cases} y_i - y_{\arg\min\limits_{\substack{j\ s.t.\\ t_i \neq t_j}} d(x_i,x_j)}, & t = 1 \\ y_{\arg\min\limits_{\substack{j\ s.t\\ t_i \neq t_j}} d(x_i,x_j)} - y_i, & t_i = 0 \end{cases}$$
  The metric that we used is Euclidean distance on full covariates. It is possible due to the small dimension, and we preferred that metric because it is often stronger in the sense of less variance. As mentioned earlier, we applied normalization before fitting the model.

We expect the methods' ATE estimations to be consistent with one another.

Confidence Intervals

For each model's ATE estimation, we calculated a confidence interval by sampling 100 samples of size 50000 (like the full dataset size) from the empiric distribution of the dataset and using the percentile bootstrap CI method.

Sensitivity Analysis

Using "method 1" from lesson 11 we calculated the tipping point λ^* for each of the four methods discussed earlier, using the following formula:

$$Bias_1 \coloneqq \lambda \cdot \frac{1}{n}\sum_{i=1}^{n} \widehat{\sigma_1}(x) \cdot (1-\hat{e}_i), \widehat{\sigma_1}(x) \coloneqq \sqrt{\widehat{\mathbb{V}ar}[Y_1|\, x_i]} = \sqrt{\widehat{\mathbb{E}}[Y_1|\, x_i] - \widehat{\mathbb{E}}[Y_1|\, x_i]^2}$$

$$Bias_0 \coloneqq \lambda \cdot \frac{1}{n}\sum_{i=1}^{n} \widehat{\sigma_0}(x) \cdot \hat{e}_i, \widehat{\sigma_0}(x) \coloneqq \sqrt{\widehat{\mathbb{V}ar}[Y_0|\, x_i]} = \sqrt{\widehat{\mathbb{E}}[Y_0|\, x_i] - \widehat{\mathbb{E}}[Y_0|\, x_i]^2}$$

$$\lambda^* = \lambda \; s.t. \; ATE + (Bias_1 + Bias_0) = 0$$

$$i.e. \lambda^* = \frac{-ATE}{\frac{1}{n}\sum_{i=1}^{n} \widehat{\sigma_1}(x) \cdot (1-\hat{e}_i) + \frac{1}{n}\sum_{i=1}^{n} \widehat{\sigma_0}(x) \cdot \hat{e}_i, \widehat{\sigma_0}(x)}$$

where $\hat{e}_i = \hat{P}(t_i = 1|x_i)$ is the propensity score, estimated using logistic regression, and $\widehat{\mathbb{E}}[Y_1|\, x_i], \widehat{\mathbb{E}}[Y_0|\, x_i]$ are estimated using logistic regression.

Model Selection

We performed 3-fold-cross-validation to choose the best model for each of our methods. For the models of S-learner and T-learner ($f(X,T)$ of S-learner and $f_1(X)$ and $f_0(X)$ of T-learner), we tested logistic regression, decision tree, gradient boosting, random forest, and gaussian NB. For estimating propensity (for IPW) we tested only logistic regression and gaussian NB since this method requires a probabilistic classifier. For each task we chose the model which achieved the highest accuracy.

The cross-validation results are summarized in the following table:

	Logistic Regression	**Decision Tree**	**Gradient Boosting**	**Random Forest**	**Gaussian NB**
S-Learner	0.605	0.537	**0.613**	0.603	0.565
T-Learner f_1	0.606	0.533	**0.615**	0.603	0.563
T-Learner f_0	0.605	0.531	**0.616**	0.603	0.568
Propensity	**0.548**	-	-	-	0.539

Therefore, we chose logistic regression for estimating propensity and gradient boosting for the tasks of S-learner and T-learner.

Results & Analysis

To test the four different methods and their estimations, we calculated the ATE index. For each method, the ATE was first calculated given the full dataset, and second on multiple samples creating confidence intervals. The results are summarized in the following table.

Method	ATE Value	Confidence Interval
IPW	0.001	$[-0.01, 0.008]$
S-Learner	0	$[-0.002, 0.001]$
T-Learner	-0.022	$[-0.039, 0.007]$
Matching	-0.026	$[-0.042, -0.005]$

Results are illustrated in *figure 3*.

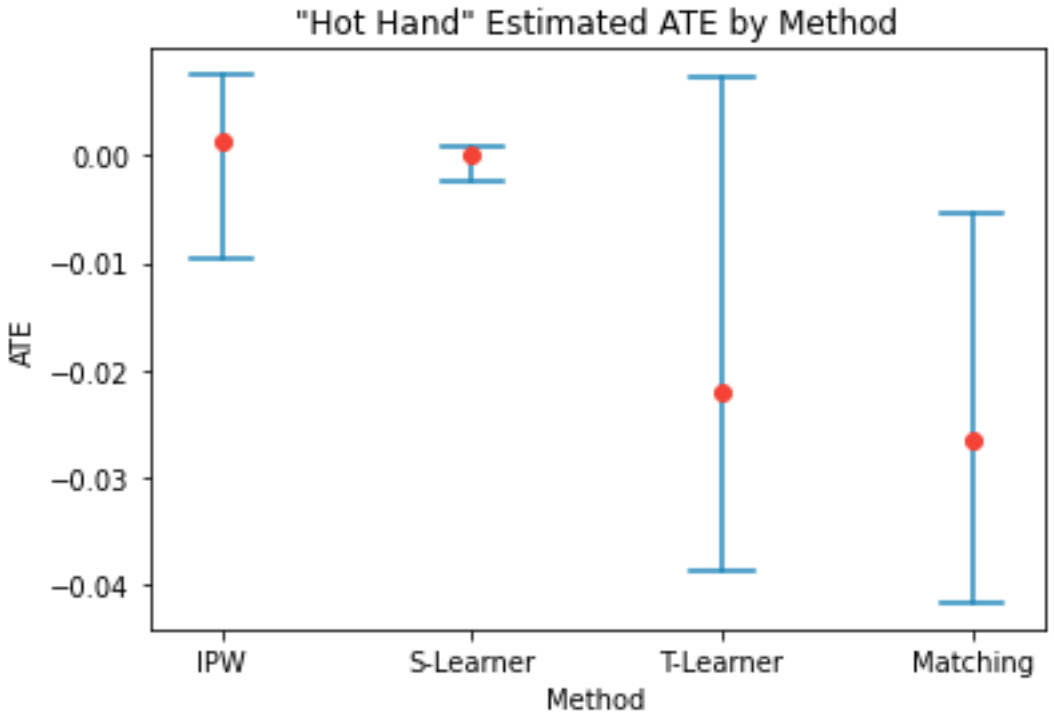

Figure 3 – ATE estimation by each method, with confidence intervals using the percentile bootstrap CI method

We notice that all the methods estimated an ATE value which is around 0, and that all the confidence intervals include 0, except of the confidence interval of matching which is negative. Furthermore, all the methods estimations are close to each other, which shows the consistency between all methods. The confidence interval of S-learner is narrower than the other confidence intervals, possibly because S-learner can be biased toward zero.

Since most methods' confidence intervals include 0, we can say that we didn't find significant evidence of a "hot hand" effect in the data tested.

Sensitivity Analysis

For method ATE estimation, we estimated the "tipping point". The results are summarized in the following table.

Method	ATE Value	Tipping Point
IPW	0.001	-0.003
S-Learner	0	0
T-Learner	-0.022	0.046
Matching	-0.026	0.055

As seen in the table, the tipping point values are very close to zero. Therefore, we conclude that the sensitivities of our models are very high, meaning that even a low level of unmeasured confounders would change the sign of the ATE result (e.g., positive instead of negative). Also, these results support the insignificancy of the evidence of a "hot hand" effect in the data tested.

Discussion

As described in the results & analysis section, we didn't find significant evidence of a "hot hand" effect in the data tested. Hence, we do not reject the null hypothesis and conclude that "hot hand" does not exist in the NBA league. However, if the matching method is more accurate than all other methods and the other methods are relatively highly biased then there is a negative average treatment effect.

Later, we present some possible weaknesses of our project methods which might affect this conclusion.

Further research may include CATE estimation and ATE estimation for other sports fields.

Possible Weaknesses

Our project suffers from some weaknesses, due to limited computational power, missing data, cases in which our assumptions have not met, and known weaknesses of the different learners.

The list of possible weaknesses is as follows:

- Missing feature in the model – which team was the last to score.
- The SUTVA assumption has a limited level of certainty, as mentioned before.
- The existence of unknown confounders might undermine the ignorability assumption – as discussed in the sensitivity analysis subsection.
- Measuring ATE using IPW – one problem with IPW, as discussed in class, is that estimating propensity is not accurate and its weighting by inverse can create large variance and large errors.
- The gradient boosting models of S-learner and T-learner might also be inaccurate.
- S-learner might vanish the effect of the treatment feature by giving a small weight to the feature of S, while T-learner may have a greater bias.

Bibliography

https://en.wikipedia.org/wiki/Hot_hand

[1] Gilovich, Thomas, Robert Vallone, and Amos Tversky. "The hot hand in basketball: On the misperception of random sequences." Cognitive psychology 17.3 (1985): 295-314.

[2] Kahneman, D. (2011): Thinking, Fast and Slow, Farrar, Straus and Giroux.

[3] Thaler, R. H. and C. R. Sunstein (2008): Nudge: Improving Decisions About Health, Wealth, and Happiness, Yale University Press.

[4] Miller, Joshua B., and Adam Sanjurjo. "Surprised by the hot hand fallacy? A truth in the law of small numbers." Econometrica 86.6 (2018): 2019-2047.

[5] McNair, Brian, et al. "The Hot Hand and Its Effect on the NBA." arXiv preprint arXiv:2010.15943 (2020).

[6] Avugos, S., Köppen, J., Czienskowski, U., Raab, M., & Bar-Eli, M. The “hot hand” reconsidered: A meta-analytic approach. Psychology of Sport and Exercise, 14(1), 21-27.