The causal effect of date's attractiveness on the desire to continue for another date

Matan-el Shpiro, Noam wolf

211419098, 318556206

{shpiro, wolfnoam}@campus.technion.ac.il

March 2020

Contents

1 Introduction

This paper is based on a Speed Dating dataset, which was collected through an experiment conducted by [1]. The dataset was created by an experiment of meetings through speed dating, in which participants engage in four-minute conversations to determine whether or not they are interested in meeting each other again.

One of the findings of [1] was that men respond more to physical attractiveness while women put greater weight on the intelligence and the race of the partner. Reading [1] has led us to the main research question in this causal inference study - **What is the causal effect of date's attractiveness on the desire to continue for another date?**

This question is a causal question because it is comprised of a *treatment* - level of date's attractiveness rating, and an *outcome* - whether or not the participant is interested in another meeting. In order to make the treatment binary, we decided that participants who rated their date's attractiveness with score greater then seven (out of ten) will be considered in the treatment group. In this project, we decided to examine this effect on the entire population and on each of the genders separately, in order to find out if there exists any difference between them.

2 Data

The data consists of 4262 observations, almost the same number of men and women, that were collected through 10 speed dating events conducted over the years 2002-2004. Each event was comprised of around 10-20 participants of each gender , that were drawn from students in graduate and professional schools at Columbia University.

2.1 Data collection

At the start of each speed dating event, the host instructed the participant to seat at two-person tables, when the females were told to seat in one side of the table and the males across from them. The males were instructed to rotate from table to table so that by the end of the dating session they had rotated to all of the tables, meeting all of the females. After any four minutes speed date, all the participant filled a scoreboard for the person with whom they were just speaking. The scoreboard was consist of a yes/no answer which indicates whether they would like to date the other person again, a list of six attributes on which the participant was to rate his/her date: Attractiveness, Sincere, Intelligent, Fun, Ambitious, Shared interests, and some more questions about the date.

2.2 Data covariates

The following tables present all of the non post-treatment covariates (and some post-treatment covariates we want to mention).

Participant personal features			
Name	**Description**	**Type**	**Values**
gender	gender of the participant	Binary	0 - Female 1 - Male
age	age of participant	Ordinal	age of participant
field	field of study	String	field of study
field_cd	field coded	Categorial	1-18 - coded field of study, when 1-17 are fields like Law, Math, Social Science etc. and 18 means "other"
race	race of participant	Categorial	1 - Black/African American 2 - European/Caucasian American 3 - Latino/Hispanic American 4 - Asian/Pacific Islander/Asian American 5 - Native American 6 - Other
from	where the participant is originally from, before coming to Columbia	String	origin state of participant
income	income of participant	Ordinal	income of participant
sports, tvsports, exercise, dining, museums, art, hiking, gaming, clubbing, reading, tv, theater, movies, concerts, music, shopping, yoga	rating of participant's interest in activity	Ordinal	1-10 - scale of interest in activity
round	number of people that met in event	Ordinal	number of people that met in event
position	station number where met partner	Ordinal	station number where met partner
position1	station number where started	Ordinal	station number where started
order	the number of date that night when met partner	Ordinal	number of date that night when met partner
partner	partner's id number the night of event	Ordinal	number of date that night when met partner
pid	partner's iid number	Integer/Ordinal	partner's iid number
match	Indicates whether both of the participants were interested to go on another date	Binary	0 - No 1 - Yes

Questions about partner and dating			
Name	**Description**	**Type**	**Values**
imprace	importance of participant that a person he/she dates with, will be of the same racial/ethnic background	Ordinal	1-10 - level of importance
imprelig	importance of participant that a person he/she dates with, will be of the same religious background	Ordinal	1-10 - level of importance
goal	participant's primary goal in participating in this event	Categorial	1 - Seemed like a fun night out 2 - To meet new people 3 - To get a date 4 - Looking for a series relationship 5 - To say I did it 6 - other
date	participant's frequently go on dates	Categorial	1 - Several times a week 2 - Twice a week 3 - Once a week 4 - Twice a month 5 - Once a month 6 - Several times a year 7 - Almost never
go out	participant's frequently go out (not necessarily on dates)	Categorial	1 - Several times a week 2 - Twice a week 3 - Once a week 4 - Twice a month 5 - Once a month 6 - Several times a year 7 - Almost never

`attr`
`sinc`
`intel`
`fun`
`amb`
`shar`
	scoreboard's rating by participant at the night of the event, for all 6 attributes about partner - attractiveness, sincere, intelligent, fun, ambitious, shared interests	Ordinal	1-10 - scoreboard's rating for attribute

like	scoreboard's rating the amount that participant likes the date's partner	Ordinal	1-10 - scoreboard's rating for like

Partner's features			
Name	**Description**	**Type**	**Values**
age_o	age of partner	Ordinal	partner's age
race_o	race of partner	Categorial	1 - Black/African American 2 - European/Caucasian American 3 - Latino/Hispanic American 4 - Asian/Pacific Islander/Asian American 5 - Native American 6 - Other

Partner's questions about participant			
Name	**Description**	**Type**	**Values**
dec_o	partner's decision the night of event for continuing to another date	Binary	0 - No 1 - Yes
attr_o sinc_o intel_o fun_o amb_o shar_o	scoreboard's rating by partner at the night of the event, for all 6 attributes about participant - attractiveness, sincere, intelligent, fun, ambitious, shared interests	Ordinal	1-10 - scoreboard's rating for attribute

2.3 Data preparation

At first we removed all of the post treatment covariates (such as answers to questionnaires that the participants had to fill after the speed dating event or at it's end), we also removed covariates that may be post treatment (such as answers to a questionnaire at the middle of the event and the answers of the date to the scoreboard) and string covariates as they have to many values, and thus changing them to onehot vectors will create highly sparse data (most of them also have discrete versions, so the removal does not hurt the predictive power of our models). We also removed the like feature at the scoreboard (due to it explains some of the treatment effect, see figure 7) and id's (which can just cause overfit).

Afterwards, we trimmed the data by it's common support with the following procedure (which will be repeated for the entire population, men and women separately). We started by calculating the propensity score, as explained in section 3.3. Then we found the minimal and maximal propensity scores in the treated and control groups. Subsequently, we set the lower (and upper) boundary of the propensity score to be the maximum (minimum) between the two minimal (maximal) values. At the end, we removed from the data each unit with a propensity score that falls outside those boundaries. This trimming

removed 2.03% when done on the entire population dataset, 0.90% when done on the men dataset and leave all samples when done on women dataset. Figures 1, 2, 3, 4, 5 and 6 illustrate the propensity score before and after the trimming of the three groups we will focus on - the entire population, the men group and the women group.

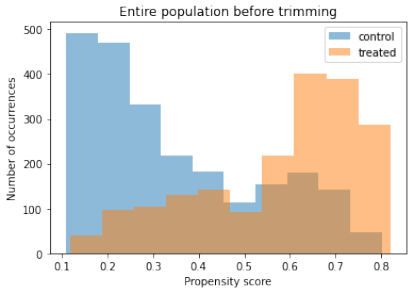

Figure 1: Entire population before trimming

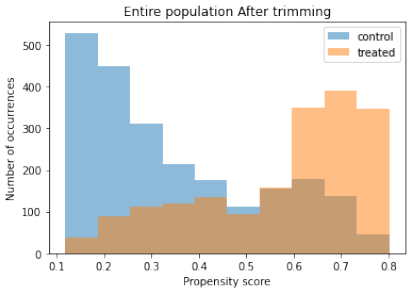

Figure 2: Entire population after trimming

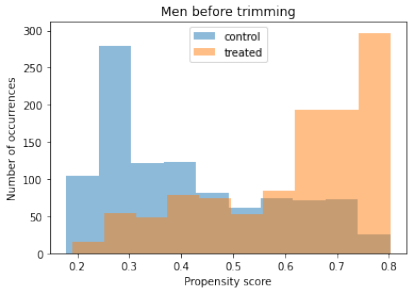

Figure 3: Men before trimming

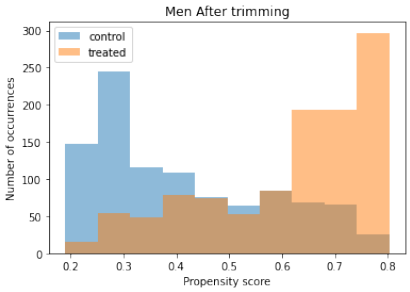

Figure 4: Men after trimming

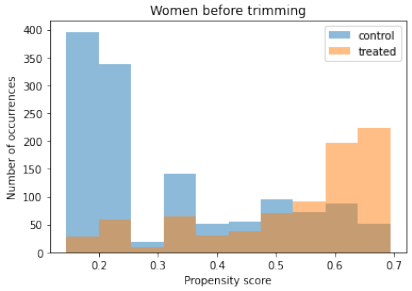

Figure 5: Women before trimming

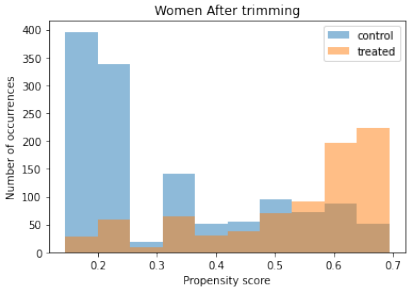

Figure 6: Women after trimming

3 Causal Inference settings and assumptions

Recall the Causal Inference assumptions we have seen in class - Stable Unit Treatment Value Assumption (SUTVA), Consistency, no unmeasured confounders (Ignorability) and common support (Overlap). We trivially assume Consistency. In this section we will examine each of this assumptions, in order to ensure their validity.

Let us denote the treatment as T, the outcome will be marked as Y and the covariates will be denoted as X. As we discussed in 1, the treatment T and the outcome Y defined as follow:

$$T = \begin{cases} 1, & \text{If the date's attractiveness rating of the participant with score greater then seven} \\ 0, & \text{Otherwise} \end{cases}$$

$$Y = \begin{cases} 1, & \text{If the participant is interested in another meeting} \\ 0, & \text{Otherwise} \end{cases}$$

3.1 SUTVA Assumption

In order to ensure the validity of SUTVA assumption, we need to examine two statements. The first one is - *The potential outcomes for any unit do not vary with the treatments assigned to other units.* This statment is valid for our data because any participant can decide "yes" (continuing to another date) or "no" as much times as he/she wants. Furthermore, we think that the influence that participant will decide "yes" if another participant decided "yes" for the same partner, is relatively small. The second assumption of SUTVA is - *For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.* This assumption holds immediately from our treatment's definition.

3.2 Ignorability Assumption

As mentioned before, the meaning of Ignorability assumption is *no unmeasured confounders*. The data probably have some hidden confounders, such as the number of dates the participant had before the speed dating event, the number of exes the participant had, and so on. We believe that our data has highly informative features, so the hidden confounders have only a slightly affect on the results. Therefore we assume ignorability for our analysis.

From the perspective of the causal graph:

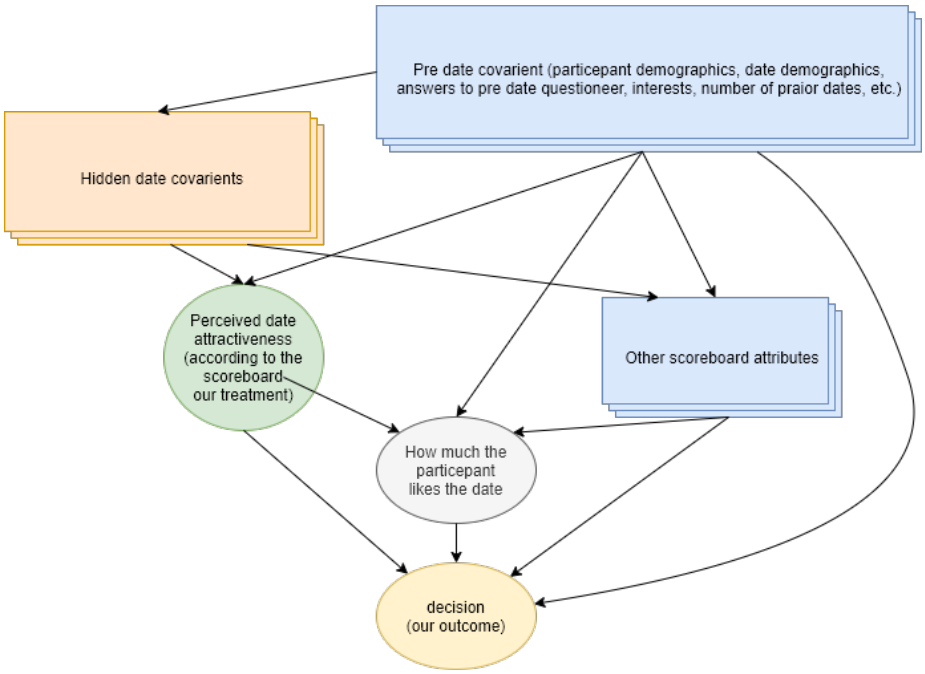

Figure 7: The causal graph

The reasoning of the causal graph 7 is the following, the pre-date covariates which represent the participant and the date demographics and character, (by interests and answers to questioners which in this graph we will assume is fully represent) and thus it effect every node. The hidden date covariates represents what happened during the date itself and the environment and thus could be affected only by the pre-date covariates. The answers to the scoreboard represent an image of the date in the participant's eye and thus (together with the demographics and character explained in the pre-date covariates) we assume that they also explain the participant's mindset. Therefore knowing them explains how much do the participant likes the date and thus we assume that the

hidden date covariates does not effect it. As with the discussion those assumptions are not correct however we believe that our data has highly informative features and thus we assume the hidden confounding effect is small. The color blue represents blocked nodes and the gray represent nodes that their features is dropped before analysis (and thus unblocked) and thus there is no blocked descendent of the treatment and all paths starting at the outcome and entering the treatment are blocked. Thus if we assume the following causal graph the back-door criterion holds.

3.3 Overlap Assumption

Recall the Overlap (Common Support) assumption from class:

$$\mathbb{P}(T = t|X = x) > 0, \forall t, x$$

In order to examine this property, we first define the propensity score - $e(x) = \mathbb{P}(T = 1|X = x)$. In order to calculate the propensity score properly, we used cross validation in the following way. We divided our data to ten groups by the speed dating event the participants took place in. For each group, we trained XG-boost model on all of the other groups and tested the model on the remaining group. This deviation was necessary in order to avoid dependency between the observations in the train group and the validation group. We used this variation of cross validation to look out for parameters that will give a high AUC measure. The AUC measure of the model we got is 0.73 for the entire population, 0.69 for the men group and 0.70 for the women group. The propensity score overlap for the entire population, the males and the females is illustrated in figures 2, 4, 6 respectively. The figures convinced us that Overlap assumption is valid.

4 Methods

In this section, we will preset the methods we used in order to estimate the Average Treatment Effect (ATE) of the three groups we focused on - the entire population, the males and the females. The usage of causal methods that we will describe are justified by the validity of the assumptions we have seen in section 3. Afterwards we will describe the way we calculated confidence intervals for the ATE using the estimators. The summary and aggregation of the results are presented in section 5.

4.1 T-Learner

We started by fitting two XG-Boost models:

$$\widehat{Y_0} \approx f_0(X), \quad \widehat{Y_1} \approx f_1(X)$$

In order to get the highest AUC measure of f_0 and f_1 we have done a grid-search on the parameters of XG-Boost as follows. To evaluate the performance of the model with certain parameters, we have done cross validation. Meaning we divided all the observations into five groups, for each group we estimated the AUC measure that the model obtain by fitting the model to all of the groups except one, and testing it on the remaining. At the end we averaged the five AUC measures and this was the measure for a certain choice of parameters.

We decided to split the *entire* population into groups, unlike we have done in the propensity score estimation, described in section 3.3. The reason for this difference is that when we tried to use cross validation on different speed dating events like in 3.3, our model preformed poorly (yielding AUC of 0.6-0.7, both for T-learner and S-learner). However using the above cross validation our models preform much better (yielding AUC of around 0.85 for the T-learners and 0.9 to the S-learners) suggesting that while our models cannot generalize for different speed dating events, they can model quite well the outcome. Thus we can estimate the counterfactual of a sample from speed dating events that where in our data. In order to be able to generalize our results we will use bootstrap confidence interval. Sampling the different speed dating events in our the data, will show as how much our conclusion will change if the data had contained different speed dating events.

Next, we calculated the ATE for the entire population, men and women:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N} f_1(x_i) - f_0(x_i)$$

Where N is the number of observations that used to calculate the ATE for each group. Models f_0, f_1 had AUC of 0.858,0.863 for the men, 0.863,0.836 for the women and 0.857, 0.849 for the entire population. The results are presented in Table 1.

	Males	Females	Entire Population
ATE	0.3201	0.2791	0.2996

Table 1: ATE estimation using T-Learner.

4.2 S-Learner

Similar to T-Learner, we started by fitting a XG-Boost model:

$$\widehat{y} \approx f(x,t)$$

The searching for good parameters of XG-Boost that will give a high AUC measure, was done in the same way we looked for parameters in the T-Learner estimation, described in section 4.1. However as opposed to the T-learner we also need to make sure our model does not ignore the treatment, so we duplicated the treatment feature a few more times. However because our models are tree based (thus looking at all the feature when doing a split and therefore are not effected by repeating the same feature) this still did not help and the ATE estimation was lower than expected. We solved this by reducing the number of feature sampled at each split, which increases the chance of splitting by the treatment (due to it's duplication) without hurting our model performance (at the price of using higher number of trees).

Afterwards, we calculated the ATE in the following way:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N} f(x_i, 1) - f(x_i, 0)$$

Where N is the number of observations that used to calculate the ATE. Model f had AUC of 0.911 for the men, 0.899 for the women and 0.906 for the entire population. The results are presented in Table 2.

	Males	Females	Entire Population
ATE	0.3227	0.2828	0.2917

Table 2: ATE estimation using S-Learner.

4.3 Inverse Probability Weighting (IPW)

In this section we will estimate the ATE by using IPW with propensity scores. As seen in class, after we estimated the propensity score $\widehat{e^i} = \mathbb{P}(t^i = 1|x^i)$, we can estimate the ATE using the following formula:

$$\widehat{ATE_{IPW}} = \frac{1}{N}\sum_{i=1}^{N} \frac{y^i t^i}{\widehat{e^i}} - \frac{y^i(1-t^i)}{1-\widehat{e^i}}$$

The results are presented in Table 3.

	Males	Females	Entire Population
ATE	0.3813	0.3244	0.3413

Table 3: ATE estimation using IPW.

4.4 Doubly Robust

As seen in class, the Doubly Robust (DR) estimator for ATE is:

$$\widehat{ATE_{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[m_1(x_i) - \frac{t^i}{\widehat{e^i}}\left(m_1(x_i) - t^i y^i\right)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[m_0(x_i) - \frac{1-t^i}{1-\widehat{e^i}}\left(m_0(x_i) - (1-t^i)y^i\right)\right]$$

Where $m_t(x) = \widehat{\mathbb{E}}[Y \mid X = x, T = t]$. We estimated $m_0(x), m_1(x)$ in two different ways. The first one is by $m_0(x) = f_0(x), m_1(x) = f_1(x)$ that we used in section 4.1. The results are shown in Table 4.

	Males	Females	Entire Population
ATE	0.3121	0.2715	0.2951

Table 4: ATE estimation using Doubly Robust with T-Learner estimators.

The second way is by $m_0(x) = f(x, 0), m_1(x) = f(x, 1)$ that we used in section 4.2. The results are shown in Table 5.

	Males	Females	Entire Population
ATE	0.3129	0.2650	0.2859

Table 5: ATE estimation using Doubly Robust with S-Learner estimators.

4.5 Confidence Interval Using Bootstrap

In this section we will present the theory of estimating a confidence interval using the pivotal (basic) bootstrap method, as described in [2]. The parameter which we interest to estimate is obviously the ATE, that we will denote as τ. At first, let as denote the ten speed dating events (which we will refer them as waves) from our data as $w_1, ..., w_n$ where $n = 10$ is the number of waves. In the methods we presented in chapter 4, we always used some function/model $g : \{w_1, ..., w_n\} \to \mathbb{R}$ in order to estimate the ATE.

We assume that the ten waves from our data - $w_1, ..., w_n$, are independent and identically distributed from the world distribution of speed dating events F. Meaning, $w_1, ..., w_n$ assumed to be realizations of random variables $W_1, .., W_n \overset{\text{iid}}{\sim} F$. Our goal is to construct a confidence interval for τ (the ATE), which is unknown parameter. By the definition of confidence interval, we are looking for boundaries $L = L(w_1, ..., w_n)$ and $U = U(w_1, ..., w_n)$ such that $\mathbb{P}(L \leq \tau \leq U) = 1 - \alpha$ for $\alpha = 0.05$.

Suppose we have a model g. We estimate τ by $\hat{\tau} = g(F_n^W)$ where F_n^W is the empirical distribution function of $W_1, ..., W_n$. The observed value of $\hat{\tau}$ is $\hat{\tau}_{\text{obs}} = g(F_n)$ where F_n is the empirical distribution function of $w_1, ..., w_n$. We

used Bootstrap Sampling (sampling $n = 10$ waves from the empirical distribution of $w_1, ..., w_n$ with replacement) in order to estimate an approximation for the distribution F_n^W. We created $m = 999$ bootstrap samples, for each bootstrap sample F_b^* we calculated $\hat{\tau}_b = g(F_b^*)$. Let us denote the random variable that presented $\hat{\tau}_b$ by $\hat{\tau}^*$.

We assume that the distribution of $\hat{\tau} - \tau$ is a pivotal quantity. Meaning:

$$\hat{\tau} - \tau \sim H \tag{1}$$

Where H do not depend on τ. We also assume that:

$$\hat{\tau}^* - \hat{\tau}_{\text{obs}} \overset{*}{\sim} H \tag{2}$$

Where $\overset{*}{\sim}$ indicates bootstrap sampling. Assumption 2 based on the premise that if F_n is close to F, the bootstrap distribution $\hat{\tau}^* - \hat{\tau}_{\text{obs}}$ will be close to that of $\hat{\tau} - \tau$. For simplification, we will ignore the fact that it is only an approximation. Let us denote the bootstrap distribution of $\hat{\tau}^*$ as B. By 2 we get $B - \hat{\tau}_{\text{obs}} \overset{*}{\sim} H$ which lead to:

$$H^{-1}(\alpha) = B^{-1}(\alpha) - \hat{\tau}_{\text{obs}} \tag{3}$$

for any $\alpha \in [0, 1]$. Under assumptions 1, 2, we have:

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\left(H^{-1}(\frac{\alpha}{2}) \leq \hat{\tau}_{\text{obs}} - \tau \leq H^{-1}(1 - \frac{\alpha}{2})\right) \\
&= \mathbb{P}\left(H^{-1}(\frac{\alpha}{2}) - \hat{\tau}_{\text{obs}} \leq -\tau \leq H^{-1}(1 - \frac{\alpha}{2}) - \hat{\tau}_{\text{obs}}\right) \\
&\overset{(*)}{=} \mathbb{P}\left(B^{-1}(\frac{\alpha}{2}) - 2\hat{\tau}_{\text{obs}} \leq -\tau \leq B^{-1}(1 - \frac{\alpha}{2}) - 2\hat{\tau}_{\text{obs}}\right) \\
&= \mathbb{P}\left(2\hat{\tau}_{\text{obs}} - B^{-1}(1 - \frac{\alpha}{2}) \leq \tau \leq 2\hat{\tau}_{\text{obs}} - B^{-1}(\frac{\alpha}{2})\right)
\end{aligned}$$

Equation $(*)$ follows from equation 3.

At the end, we can see that all we need to do, is to calculate the the respective quantities of the bootstrap distribution B we sampled, marked as $q_{\frac{\alpha}{2}} = B^{-1}(\frac{\alpha}{2}), q_{1-\frac{\alpha}{2}} = B^{-1}(1 - \frac{\alpha}{2})$, and the confidence interval will be:

$$\left[2\hat{\tau}_{\text{obs}} - q_{1-\frac{\alpha}{2}}, \quad 2\hat{\tau}_{\text{obs}} - q_{\frac{\alpha}{2}}\right]$$

5 Results

Table 6 summarizes the estimations of the ATE and Confidence Intervals (CI) from all the methods in section 4.

	Men		Women		Entire population	
Method	ATE	CI	ATE	CI	ATE	CI
T-Learner	0.3202	[0.2479, 0.3754]	0.2791	[0.2277, 0.3113]	0.2996	[0.2525, 0.3332]
S-Learner	0.3228	[0.2498, 0.3799]	0.273	[0.2265, 0.3026]	0.2998	[0.2427, 0.3230]
IPW	0.3813	[0.3248, 0.4518]	0.3244	[0.2982, 0.3612]	0.3413	[0.3078, 0.3776]
DR1	0.3122	[0.2415, 0.3690]	0.2715	[0.2203, 0.3044]	0.2952	[0.2469, 0.3290]
DR2	0.3129	[0.2403, 0.3725]	0.2650	[0.2160, 0.2949]	0.2859	[0.2377, 0.3165]

Table 6: Summary of all the results from Section 4.

6 Possible Weaknesses

During the project we had uncertainty about some of the covariates and we noticed some possible weaknesses we had. At first, the causal graph we presented in figure 7, might have some missing arrows between nodes which represent implications of covariates on others. For example, perhaps the hidden date covariates has a direct implication on the outcome (the decision) that not explained by the observed covariates (like not explaining fully the mood of the participant). Next, it is possible that the first judgment of some participants is the attractiveness of the date, which makes the other scoreboard covariates to be post-treatment and we can not use them. Another possibility is if the fact that a participant (or group of participants) rated some date with high (or low) attractiveness has a significant implication on the possibility that other participants will rated high (or low) this date, the first assumption of SUTVA (which we discussed in section 3.1) does not hold. Additionally since the ATE estimates of the doubly robust are closer to the learners from the IPW and thus our propensity score may be biased which may effect the overlap.
Another weakness is that our models cannot generalize to people that was not in the training data and thus all the generalization is dependent on the bootstrap confidence interval. Moreover we had only 10 speed dating event to construct the bootstrap confidence interval from (which might be to little) and 999 replicates (which might be to little too). Lastly our data is not i.i.d in the dates instances, but in the dating events of which we have only 10.

7 Conclusions and Discussion

In all the methods we used, this is clear that the confidence intervals of the ATE are far from including zero. Thus we reject the null hypothesis that the

ATE is zero, and deduce the ATE is positive (for both genders and the entire population). As we presented in section 5, the estimations of T-Learner and S-Learner were close in each group. The estimations of the Doubly Robust methods were similar to the original learners, and close to each other. The IPW method provides higher ATE for all the groups. Because the Doubly Robust estimations were close to the learners estimators and far from the IPW estimator, we believe that our IPW estimator is biased. All the estimators agreed on around 0.05 difference of the ATE on the man and on the woman (relatively, it is a growth of 16%), which is smaller then our primary thought. That is because 51.73% of the men in our dataset rated women attractiveness above seven, when only 38.18% of the women rated men attractiveness above seven. Therefore, what we actually found is the different meaning of rating above seven for men and women,which making the true difference in the effect of attractiveness to be quite larger.

8 Future Work

- Major issue we had to deal with was that the observations on each speed dating event were dependent between them. Moreover, the number of speed dating events was only ten. In future, we would like to work with a bigger dataset which has more independency between observations, or large number of groups that observations inside them are dependent (like in the speed dating events).

- Using more advances technique that does not require as precise learners and thus not using only the bootstrap confidence intervals for generalization.

- Setting different thresholds for the men and women, according to the quantile of treated to get the actual difference in effects of attractiveness between the genders (or using each rating as a treatment).

- Fixing our confidence interval of men VS. women ATE: during our work we tried to implement a confidence interval for the difference between men and women ATE by matching the men and women bootstrap ATE. But by mistake we matched samples with different wave selection, and therefore we did not use it. Since running the bootstrap process takes long time we chose to keep it as future work.

References

[1] R. Fisman, S. S. Iyengar, E. Kamenica, *et al.*, "Gender differences in mate selection: Evidence from a speed dating experiment," *The Quarterly Journal of Economics*, pp. 673–697, 2006, https://faculty.chicagobooth.edu/emir.kamenica/documents/genderDifferences.pdf.

[2] R. J. Tibshirani, "Bootstrap confidence intervals," *Department of Statistics, Stanford University*, pp. 1–3, 1984, https://apps.dtic.mil/dtic/tr/fulltext/u2/a147572.pdf.