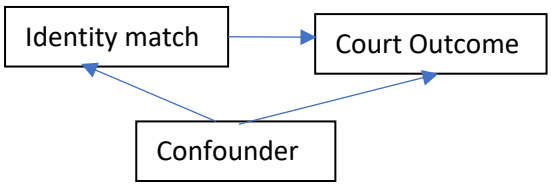
Final Assignment BRJ

Exercise 1

a) The paper considers the causal effect of gender and religion on the outcomes of court cases in India. It asks the question whether judges deliver more favourable treatment to defendants that match their identities. The causal graph that corresponds to this situation is depicted below. There, identity is characterized by the gender or religion (and caste is also researched, although the results are a bit less strong due to name matching problems). In this specific example, age of the judge could be a potential confounder. Income might be a confounder when judges are not randomly assigned. We could think of a case where wealthy individuals might be able to exert influence on which court and judge is appointed for their case, influencing the probability of an identity match. Moreover, wealth might have a direct influence on the court outcome. This could be positive, since higher fines might be given (or in the case of discrete: the probability of a fine might be higher) since wealthy individuals can easily afford them. But it could also be negative because they might have better lawyers and therefore have lower probability of being convicted.



b) The purpose of these balance equations is to "test" whether we can be confident that the randomization of the judge assignment is likely to be valid. This is important for the research design, because it gives us a sense whether our identification strategy will be successful in constructing the desired causal effect estimates. The empirical strategy is only able to identify the causal effect when we do not have to worry about hidden confounders. When randomization is done properly, we need not worry about this and hence we can interpret the coefficients from equations (1) and (2) as causal.
Basically, the equations aim to explain gender (or religion) of the judge as a function of the gender (or religion) of the defendant (and some fixed effects and controls to adjust for some effects such as location, time). When randomization is done properly, we expect the effect of the defendants identity on the judge identity to be zero (there will be no correlation between the identity of the judge and the identity of the defendant).
In the table, we can recognize the coefficients corresponding to beta_1 (= 0.0005 for specification 1, =0.0012 for specification 2), beta_2 (= -0.0007 for specification 1, = -0.0004 for specification 2), gamma_1 (=0.0001 for specification 3, = -0.0001 for specification 4) and gamma_2 (= 0.0001 for specification 3, =-0.00034 for specification 4) in equations (3) and (4). For each of the equations, there are two specifications, corresponding to the case with court-month fixed effects or court-year fixed effects. The numbers between brackets give the standard errors. The first two specification correspond to equation 3, and the second two specifications correspond to equation 4. Two of the specifications, hence, look at the randomization on the religion identity, while the other two look at the gender identity metric.
From the results, we can conclude that the randomization has probably worked well. All

coefficients (except for 2) are not significantly different from zero. Since beta_1 is significantly different from zero (and also in absolute value a bit away from zero), it is chosen to be added as a control to account for potential confounding.

c) Colliders are characterized by being affected by the treatment (identity) and the outcome (court case outcome). Firstly, there is probably no effect of identity on the type of crime that is committed (seems logical to me, when cases are randomly assigned to judges). Hence there is no arrow going from identity to the criminal case fixed effect. Moreover, to me it seems more likely that there is an effect from the fixed effect onto the outcome and not the other way around. Hence if there is an arrow between outcome and the fixed effect, I would argue that its direction is from fixed effect to outcome, and not the other way around.

d) Since we are dealing with high-dimensional data, normal propensity score matching or adjusting for covariates with linear models might not work (if the dimensionality of the data is truly larger than the number of observations). Hence we can resort to Double Machine learning. When we want to preserve the linearity of the method (like the models are implemented in the paper), we can use Double Lasso. However, a more flexible approach to capture non-linear relationships between the control variables and the outcome/treatment can be used using the general Double Machine Learning approach. In that approach, we would fit an outcome model on the judge identity measure (gender or religion) with all control variables we have available for the cases, and a treatment model on the defendant identity. And then create residualized treatment and outcome variables by using predictions from these models. Those can then be regressed on each other (residualized judge identity on residualized defendant identity) to learn the causal parameter (we will also do cross fitting and average over the two causal estimates we get from that to prevent overfitting).

Exercise 2

```
# -*- coding: utf-8 -*-
"""
Created on Wed Dec 21 10:36:06 2022

@author: joppe
"""



import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import balanced_accuracy_score

np.random.seed(12345)

X = np.random.normal(0, 1, size = (1100, 5))
y = np.append(np.ones(1000), np.zeros(100))
```

```
# The data is slightly unbalanced, so let's sample some extra ones to make the
data balanced

y0 = y[y==0]
y1 = y[y==1]
X0 = X[y==0, :]
X1 = X[y==1, :]

y0_appended = y0
X0_appended = X0

# We add as many points to y0 such that the data is perfectly balanced (50/50)
for j in range(len(y1)-len(y0)):
    index = np.random.randint(0, len(y0))
    y0_appended = np.append(y0_appended, y0[index])
    X0_appended = np.append(X0_appended, X0[index,:].reshape(1,-1), axis = 0)

# Create the complete datasets
X_appended = np.append(X0_appended, X1, axis = 0)
y_appended = np.append(y0_appended, y1, axis = 0)

# Split (ONLY AFTER DOING THE OVERSAMPLING!!)
X_train, X_test, y_train, y_test = train_test_split(X_appended, y_appended,
test_size = 0.5, random_state = 12)

model = RandomForestClassifier(n_estimators = 10)
model.fit(X_train, y_train)

# Print the accuracy (will be high!)
print("Accuracy with incorrect splitting: ")
print(balanced_accuracy_score(y_test, model.predict(X_test)))


########### With the proper split

# Now we first do the sample splitting and only afterwards the oversampling
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.5,
random_state = 12)

y0_train = y_train[y_train == 0]
y1_train = y_train[y_train == 1]
X0_train = X_train[y_train == 0, :]
X1_train = X_train[y_train == 1, :]

y0_train_appended = y0_train
X0_train_appended = X0_train

for j in range(len(y1_train)-len(y0_train)):
    index = np.random.randint(0, len(y0_train))
```

```
    y0_train_appended = np.append(y0_train_appended, y0_train[index])
    X0_train_appended = np.append(X0_train_appended, X0_train[index,
:].reshape(1,-1), axis = 0)


X_train_appended = np.append(X0_train_appended, X1_train, axis = 0)
y_train_appended = np.append(y0_train_appended, y1_train, axis = 0)

model2 = RandomForestClassifier(n_estimators = 10)
model2.fit(X_train_appended, y_train_appended)

print("Accuracy with correct splitting: ")
print(balanced_accuracy_score(y_test, model2.predict(X_test)))
```

The output from this is:

Accuracy with incorrect splitting:

0.919150641025641

Accuracy with correct splitting:

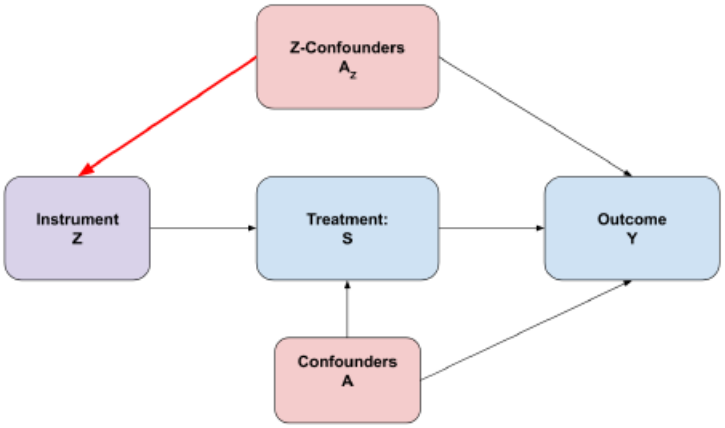
0.47035573122529645

Exercise 3

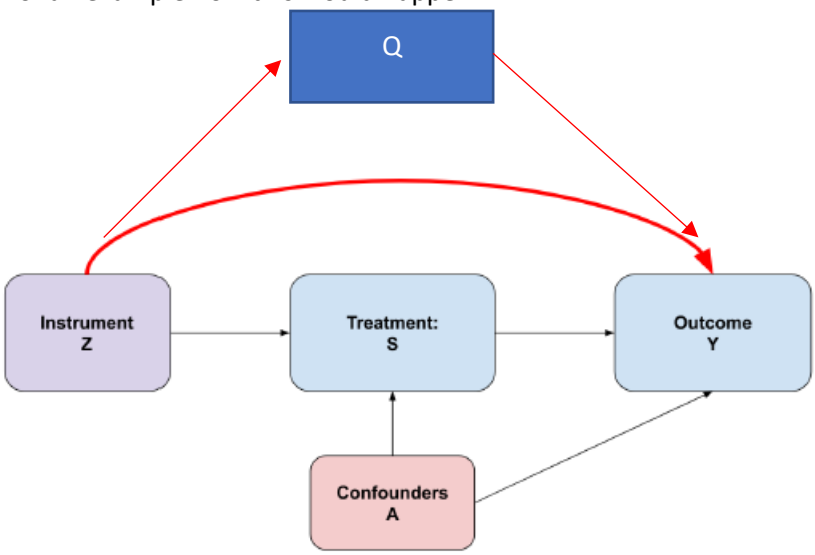
a) The outcome is health behaviour (as measured by individual's degree of mobility and shopping behaviour, and also covid infections/death rates), the treatment is Fox News Viewership (measured by the Nielsen 2020 ratings) and the instrument is the rank of the Fox News channel in the channel list (which is exogenously determined).

b) The three assumptions are:

1. Exogeneity: There are no Z-confounders. This means that there are no unobserved(!) variables that affect both the instrument and the outcome (in the graph: there would need to be an arrow from Z to the instrument and an arrow into the outcome). For violation of this assumption, we would need to find a variable that influences the rank of the channel, as well as health behaviour. In the paper, some variables X are controlled for in the first and second stage. Hence, we would need to find a variable that could violate the (conditional) exogeneity of the rank of the channel that is not in X. Judging from the list of variables that are controlled for in the regressions (both in 2SLS and reduced form), it seems that many potential variables that could create this violation are already controlled for (many demographic, political dimensions).



2. Exclusion: which means that the outcome is only affected by the instrument through the treatment. When this would be violated, this means there would be an effect from the rank of the FNC on another variable than viewership that influences health behaviour (or a direct effect). In the graph this would be depicted by the red arrow, or the red arrow going into another variable Q that affects Y). This seems very unlikely in my opinion and I cannot really think of an example how this would happen.



3. Relevance: the instrument should be sufficiently correlated to the treatment variables. In the graph above, this basically means that there should be an arrow going from the instrument to the treatment. This can be "tested" by looking at the F statistic of the first stage regression. Note that this is only an informal test of correlation and cannot say that the instrument has a causal relationship to the treatment. This should follow from domain knowledge. A violation of this assumption in this case would be that the rank of FNC does not have an effect on viewership. This makes intuitive sense as people often start zapping at low numbers and spend more time at channels early in the lineup.

4. (Monotonicity). A last but less discussed assumption is the monotonicity of the instrument. This means that there is a one-directional effect of the instrument on the treatment. For this situation, this could be violated, if viewership does not monotonically

decrease as the rank of the channel goes up. A case could be made for this, since there are many anchor channels from which people might start zapping (think of HD channels starting at 101 instead of non-HD channels starting at 1).

c) I think that the overall idea behind the research design is good and intuitive. However, there are multiple smaller points which make me less enthusiastic about this paper. Firstly, an IV approach only identifies the causal effect for compliers (local average treatment effect). Since the treatment and instrument are continuous (or count) data, compliers are a bit more complicated to define. However, I would interpret it as the people that only watch more(!) Fox News, simply because the rank is lower. This of course does not invalidate the research design, but it does significantly change the potential interpretation of the results. When this group of people is very small, the societal impact of the results is not very interesting. It would be a great improvement when the paper discusses this a bit more explicitly and sheds a bit more light on how the results should be considered. In the abstract, for example, the causal phrasing might suggest that the effect goes beyond this potentially small group of compliers. Moreover, (as I described above), another important assumption of an IV analysis is monotonicity. I think a strong case can be made that monotonicity is violated. With modern digital TV subscriptions, there is usually a plethora of "anchor channels". For example, the HD channels might only start at 101, and they might have higher viewership than the equivalent non-HD channels starting at 1. When these effects are strong enough, this might invalidate the approach.
Lastly, reverse causality between viewership and the rank of FNC could be a serious issue for the research design. When the viewership of a channel influences its position in the programming, this will break the IV design. Maybe previous literature talks about this issue (my bad, then), but otherwise, I feel like the authors could have elaborated upon this more. This effect (of viewership on the channel list) would then have to come directly from the viewership and not other variables (like demographics), since the exogeneity checks should rule out the indirect reverse causality from viewership through those variables.
Nevertheless, I think the main idea of the paper is very strong and nice. When discussing (and verifying) some of these potential shortcomings a bit more closely, I think the paper could become even better!

Exercise 4

1. Payoff does not depend on anything else besides the predictions (generalization of benefits of convincing a voter are homogeneous). This means that the decision to run a video is only dependent on the potential of persuading a voter (which the model hopefully predicts well). However, there might be other reasons that would be important for the decision to launch a video. An example in this context could be the preferences of the candidate. Maybe he/she is more comfortable with a certain message and finds a video that might persuade people as not fitting to their personality. It could also be that the value attached to persuading a voter is not homogeneous. A candidate might definitely want to persuade a subgroup of voters, while not caring about another one, and hence the benefits of convincing someone are not evenly distributed across voters.
2. Actors do not change behaviour in response to the prediction algorithm. When we would deploy this prediction algorithm and use it the predictions should reflect the true opinion of people watching the videos. This could be violated as soon as multiple videos have been run already and voters have made up their mind (maybe partly because of the videos). The

model has then (potentially) been trained on people with an open mind. The voters, however, may not react anymore to the videos as the model would expect. This would constitute to a domain shift of voters and should be taken into account by the model.

3. People respond predictably to the algorithm (decision makers follow the algorithm threshold rule). The people deciding on whether to run a video after predictions should comply with the predictions. In the following situation this could be violated. Maybe there is a political veteran that would not follow the recommendations by the model, because he/she feels they have a better understanding of voter responsiveness.
4. Political advisers (decision makers for the video) get feedback on prediction accuracy (to assess domain shift). When the decision makers would get input on whether the actual voter persuasion based on a launched video is in line with what the model thinks, this will improve the trust in the algorithm. To do this, surveys could be done among voters. However, the final results of an election are only available after the election, so this could violate this feedback requirement. Even though you can do surveys to get feedback, you can never get the full picture of the effect until after the election. And even then, it will be hard to pinpoint whether people were persuaded by the video alone, and not by other factors. Then the people making decisions do not know how their decisions are working out in reality and therefore might have less confidence in the predictions.

Exercise 5:

1. Payoff does not depend on anything else besides the predictions (benefits of admitting an applicant from a college essay are homogeneous) . In the context of college admission essays, this means that the value of admitting an arbitrary applicant are homogeneous and there is not more or less payoff for specific candidates. This condition depends on the values of the decision makers. When every person is valued completely equal, given the score, this will probably be satisfied. However, with a lot of university valuing diversity, there might be also other factors that might make the benefit heterogeneous.
2. Actors do not change behaviour in response to the prediction algorithm. This means that applicants will not actively try to game the system. In class we talked about the case where you would add the words “Oxford” or “Cambridge” in white (or invisible) text to get higher scores on job interview scores. Something similar could be done for college admissions, such as the names of some fancy private prep schools.
3. People respond predictably to the algorithm (decision makers follow the algorithm threshold). For this example, this would entail that when the algorithm provides a list or score for the candidates, these recommendations should be followed by the admissions officers. When the process is fully automated, this condition is probably satisfied, since humans have no discretion in the procedure.
4. Admissions officers get feedback on prediction accuracy (to assess domain shift). The model will have been trained on previous admissions (and hopefully successfully, but since for rejected people we do not know if they would have succeeded, this could be challenging!). However, requirements for successful college performance might change. Therefore, admission officers should get feedback on whether the predictions were right or not. This could be hard, since outcomes (like grades, or other measures of success) will generally only be available after students are studying there for a while.

When the algorithm is only used for advising, the first criterion can be made less stringent. When the school would like to do some positive discrimination (affirmative action), the benefits of accepting one candidate over another might not be homogeneous. However, after an initial screening, the

humans could still intervene to make optimal decisions under the utility function of the school. The admission score for the applicants will then just be an input to the final decision. The second criterion would still be important. If applicants change their behaviour, they could still influence the decisions without the decision makers noticing, which would be suboptimal. The would basically get a score by the algorithm that is too high. If the admissions officers would notice this, they could correct it, but if it would go unnoticed, this criterion is still very important to satisfy. The third criterion also changes by definition, since the discretion of acceptance is given to the application officers and not to the algorithm. The last criterion (feedback) would still be important, since the model should be used to inform the final decision made by the decision maker. Without the accuracy assessment, there might not be much value in presenting the recommendations by the algorithm at all.

Exercise 6

a) The article in essence refers to statistical parity. It would want to see that outcomes are the same when people with the same characteristics, except for something like a name (or other protected feature). Suppose we have two groups A ("white names") and B ("non-white names"). When these groups have the exact same characteristics, except for the name, we would like to see the same outcome: call back rates (which is our outcome variable Y, with corresponding predictions Y_hat). However, from this quote it seems that this is not the case. "Marianne Bertrand, an economist at the University of Chicago, and I conducted the first study: We responded to actual job listings with fictitious résumés, half of which were randomly assigned a distinctively black name. The study was: "Are Emily and Greg more employable than Lakisha and Jamal?" The answer: Yes, and by a lot. Simply having a white name increase callbacks for job interviews by 50 percent."
b) There are two more notions of statistical fairness: separation (e.g. equalizing recall across groups) and calibration (equalizing precision across groups). With the statistical parity, we might disadvantage certain groups in terms of accuracy if there are different base rates. The following two notions do allow for some divergence in metrics.
    a. Separation. This would require that the error metrics should be equal conditional on the true label. This would allow for different base rates in the callback rates for "white" and "non-white" applicants, but would equalize the true positive rate (recall for positive class) and the false positive rates between the two groups. Hence it would say that the predicted probability of a callback is the same for the two groups given that the person in reality received a callback.
    b. Calibration (or also called sufficiency). This concept also requires equalization of error rates, but then conditional on the predictions (it equalizes precision for the positive and negative outcomes). Instead of conditioning on the outcome (like the separation case), it conditions on the prediction and equalizes the probability of getting a callback, given that the model predicts that the applicant gets a callback. This can be achieved by well calibrated models, which means that the modelled (predicted) probability for a certain class reflects the true probability of observing that class.

Exercise 7

a) The main difference between global and local explanations is that global explanations look at what variables are generally important for predictions, while local explanations look at

which variables are important for a specific prediction. A local explanation in this context (Shapley values, for example), would look at a prediction of corruption for a certain municipality in Brazil and explain which features contribute most to the prediction for that specific municipality. It might have a very large budget for infrastructure projects (random example), which drives the predicted corruption for that municipality.

Example of situation for local importance: suppose that the model predicts that a certain municipality has corruption. Local feature importance might give investigators leads on where to start their investigation to uncover the corruption. It might make their search easier and more efficient.

Example of situation for global importance: suppose that the government wants to implement some new laws to make corruption harder. By looking at some global feature importance metrics, they might get an idea what is important for the model to predict corruption. If those are variables that also have semantic meaning for the politicians, they might be able to use it to tailor their laws to be most effective and tackle the issues that the model is able to capture.

b) There are two very distinct goals of the procedures described in sections 4 and 5. In section 4, the main goal is to use the predicted corruption rates to augment current datasets for social science research. In that respect, the most important aspect is that the predicted corruption rates are as close to the real ones as possible (unbiased and consistent), since they should be used to draw conclusions about the potential mechanisms behind corruption. Hence you want to pick the set of features that optimally makes predictions. Section 5, however, is about policy implementation. When using predictions for policy, there might be different things that are important, next to accuracy. Maybe we cannot use some features because they are protected (like political party), or we want to get a set of features that gives better calibrated predictions. Since there is a different goal, the final set of features used for the algorithm will probably be different. The baseline model is created to achieve the best predictions in regular terms (accuracy), and hence for the task in section 4. The goals of section 5 (helping with decisions) can be broader than just having the best possible accuracy. Fairness is an important aspect, but the system should also not be easy to game by municipalities. Maybe therefore you would want to add different variables which municipalities have less discretion over to influence (then they cannot change these variables to prevent inspection). Moreover, for the second task, maybe not only the presence of corruption is important, but also the importance or scale of it. Hence using the benchmark model (for probability of corruption) is not the best model, since we would also want information about the scale or type of corruption. The model could then be extended to also capture this. Maybe in the historical corruption reports there are some measures of gravity.

c) The text refers to the fact that audit rates should be equalized across political parties. This refers to the statistical parity concept (where outcomes are equalized among groups). It is slightly different from the type of statistical parity that we considered before, since the prediction model is not changed. Instead, a post-processing step is added that is completely separate from the prediction model. This would not be possible if we would not have access to the protected attribute (in this case political allegiance). Instead, some pre-processing techniques could then be used in the model. Some simple approaches of pre-processing will, however, not result in the statistical parity that is described in the paper.