Gender and Business Corruption in India

Johannes M. Halkenhaeusser

Minerva Schools at KGI

IL181.004 Prof. Scheffler

Fall 2020

Introduction

What is the causal effect of having a female business owner on reported business corruption in India? To answer this question, this project will specify a structural causal model for corruption in the business context, estimate the propensity score using a logistic regression and using a neural network, and compare the results from both estimations when using them for propensity score matching. I find that the limitations and requirements for matching make it difficult to build an accurate propensity score model that deals with data limitations. The effect of having a female manager is insignificant across the three different propensity score models employed.

Corruption is one of India's harshest barriers to economic development. Specifically, in the business context, evasion of profit taxes and bribery of officials are anti-competitive, weaken institutions, and hold back coordination (Svensson, 2005), making it a crucial topic in development economics.

Females are associated with a lower likelihood of being corrupt. For example, women are less likely to be corrupt when in a government position (Dollar, Fisman, & Gatti, 2001). They are also associated with lower levels of income tax evasion (Cabral, Kotsogiannis, & Myles, 2019). In the business corruption literature, women have female firm ownership or management has also been found to be robustly correlated to reported bribery using cross-country analysis (Breen, Gillanders, McNulty, & Suzuki, 2017). Experimental settings have made more mixed findings with women being more risk averse and hence less likely to participate in corrupt behavior (Frank, Lambsdorff, & Boehm, 2011) or as tolerant of corruption as men (Alatas, Cameron, Chaudhuri, Erkal, & Gangadharan, 2009). However, the latter study finds differing levels of acceptance depending on the country, reporting a gender difference in Australia that does not replicate in India, Singapore, and Indonesia. With the former field experiment being conducted in Germany,

the experimental evidence with causal claim makes mixed findings. I will hence try to close this gap using observational evidence to estimate the effect of having a female business owner on the reported corruption using a comparison of three propensity score models.

Matching for Causal Inference

The fundamental problem of causal inference is that the counterfactual outcome is unknown. Therefore, any correlation that is found in observational data could be spuriously created by confounding factors. Matching solves this problem as it tries to balance the covariate factors.

Each observation in causal inference has a number of potential outcomes Y that change after being intervened on with treatment value X. Any set of variables Z that affects both outcome and treatment is considered a confounder. In the absence or independence of any confounders, the treatment effect is given by the expected difference between the two potential outcomes. In the context of the above research question, the causal effect would ask the question: What is the change in bribery if the only factor that is being changed is the gender of the business owner? The treatment effect would hence be given by following formula:

$$E(Y_{X=1} - Y_{X=0}), where\ X, Y \perp Z$$

In the language of do-calculus the causal effect can be re-expressed as:

$$P(Y \mid do(x)) = \sum_{z} \frac{P(X = x, Y = y, Z = z)}{P(X = x \mid Z = z)}$$

The "do" in do-calculus is used to indicated that X is intervened on. That is, the only changed factor is the observation's treatment value. The denominator of above equation $P(X \mid Z)$ is the probability of the treatment value given the confounding variables. By controlling for this probability, the relationship between X and Y is isolated. The probability of receiving treatment is called the propensity score.

Matching uses the covariates to create a sample where each observation in the treatment group has an observation in the control group that has similar/the same covariate values Z. This allows the creation of two sets of observations that are identical on the observed covariates with the only exception being a difference in the treatment value. A fundamental assumption is that all

covariate variables are observed and correctly specified. If they are un-observed, the variables that are being matched need to have such a strong correlation with the unobserved confounders, that they are similar for each matched pair. Propensity score matching uses the propensity score as summarizing metric of the covariate values for the observation. Using the propensity score suffices because the information of the covariates is contained within the propensity score. Unlike the equation above, the observations are not weighed by their propensity score, but the propensity score balance achieved across samples means that each group has the same average propensity score. An accurate and precise estimation of the propensity score is hence key to match the right observations.

Propensity score matching has received criticism from prominent causal inference. (King & Nielsen, 2019) find evidence that propensity score matching is strongly dependent on the model with which the propensity score is estimated and that further propensity score matching is not equipped to balance observations in the same way a blocked randomized experiment is. The danger of model induced bias is echoed by (Pearl, 2009) who warns that estimating a propensity score without carefully checking for colliders and possible confounders, will yield a propensity score that does not represent the real world. He sees propensity score matching as a method that has been misinterpreted as a one-size-fits-all approach that is applied without an understanding of the gravity of assumptions made. One such assumption is that the set of covariates used to estimate the propensity score is sufficient to d-separate (cut off all correlational flows) between the treatment and the outcome. This requires strong knowledge of the structural causal model at hand.

Collider bias is induced into a model when a variable is controlled for that is an outcome of treatment and outcome variable. If there are three variables of which two are the causes to the third and the causing variables are independent, then only conditioning on their common outcome

(the collider) renders them dependent. If one conditions on a collider when estimating the propensity score, one may actually induce an additional spurious correlation and hence introduce collider bias. The bias does not only occur when conditioning on a direct outcome of treatment and outcome variable but any collider that may block any backdoor path between treatment and outcome.

While Pearl (2009) and King & Nielson (2019) are correct that causal inference is not a fool-proof method, using a propensity score still allows us to efficiently coerce covariate information into a single value for each observation. It further can provide the necessary covariate balance when and is able to disentangle a causal effect if the correct causal model is specified. Hence, I will attempt to construct a causal model that can be used to estimate the propensity score.

DAG

The available confounders of the relationship between gender and perception of corruption that are specified in the dataset are largely exogenously given characteristics such as the region of the firm, the sector the firm operates in, size at start-up, and the type of ownership (Abdixhiku, Krasniqi, Pugh, & Hashi, 2017). I specifically exclude variables that are other firm outcome variables as I cannot exclude them from the collider list. A business with a female manager may have an overall higher profit than a business with a male manager. At the same time, profit may be driven by the fact that a business has a different perception of corruption, bribes more, and hence has higher profits. This would create a spurious correlation between treatment and outcome that biases the treatment effect.

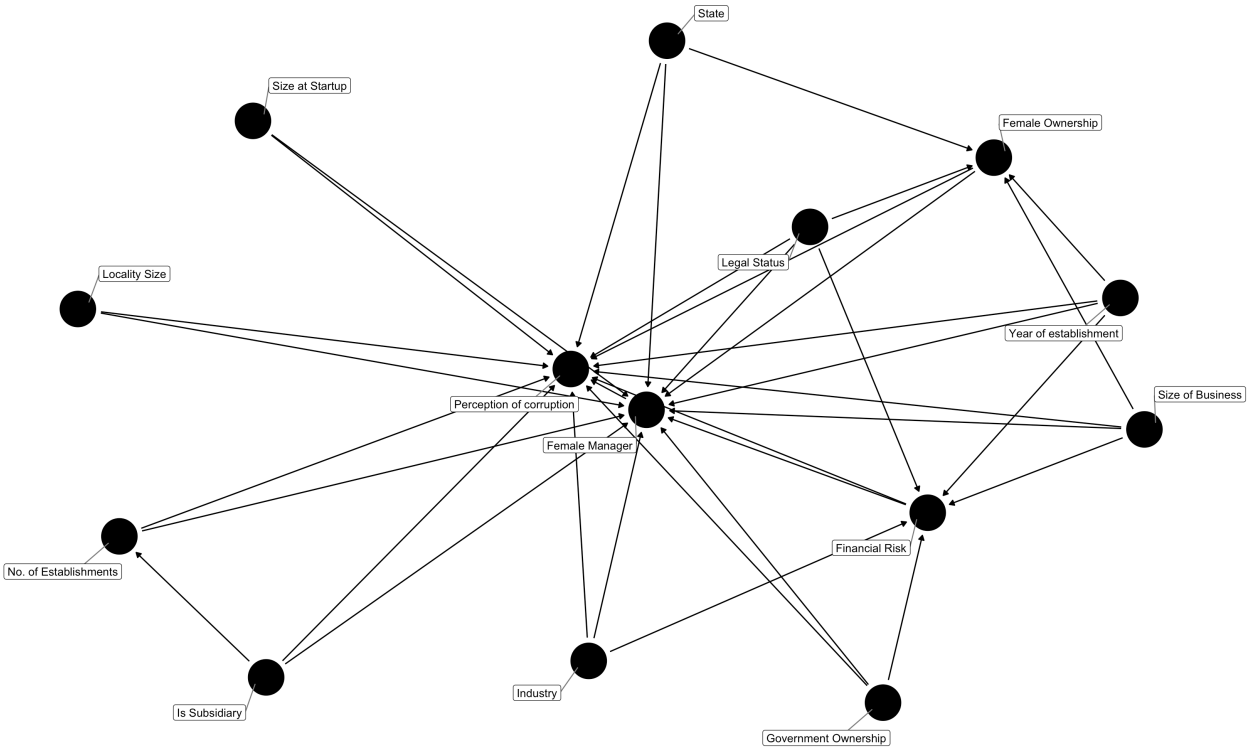


Figure 1. DAG of the causal relationships including the confounders. I have added a variety of interactions between the confounders. However, they are not exhaustive and are less relevant once matching has controlled for them. They serve as exemplary connections to show that there is added complicatedness from connections between the DAG. A larger version of this figure is in the Appendix.

Data

The World Bank's Enterprise Survey (BEEP) studies firm-level characteristics and outcomes across the world and across countries (World Bank, 2014). While it captures a multitude of outcome variables, it only captures a limited number of indicators describing the context of firm. To estimate the relationship above, I use 8451 observations of the variables estimated by the questions in Tab. 1.

To evaluate the overall distribution of the covariates and their balance, I estimate summary statistics (Tab. 2) and plot the distributions split by treatment group (Fig. 2). A common metric for the balance of the variables is to estimate the absolute standard mean difference, which is the difference between the means of the variables by treatment/control group, weighted by standard deviations to estimate the spread of the variables:

$$SDM = absolute\left(\frac{\overline{(X}_{T=1} - \bar{X}_{T=0})}{\frac{1}{2} * \sqrt{\sigma^2_{T=1} + \sigma^2_{T=0}}}\right)$$

As illustrated by the varying means in the Tab. 2 and the distributions in Fig. 2, there is a significant imbalance between the covariates. To evaluate the balance of covariates, I will continue to evaluate the SDM. Fig. 3 gives an overview of how the SDM compares across the variables.

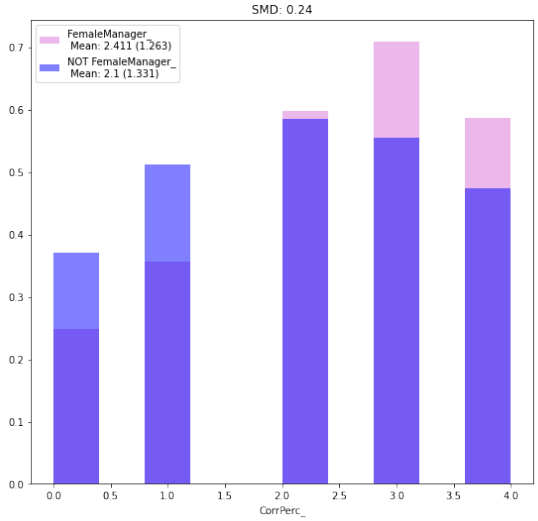
Table 1. Variables and their respective questions in the BEEP questionnaire (World Bank, 2014).

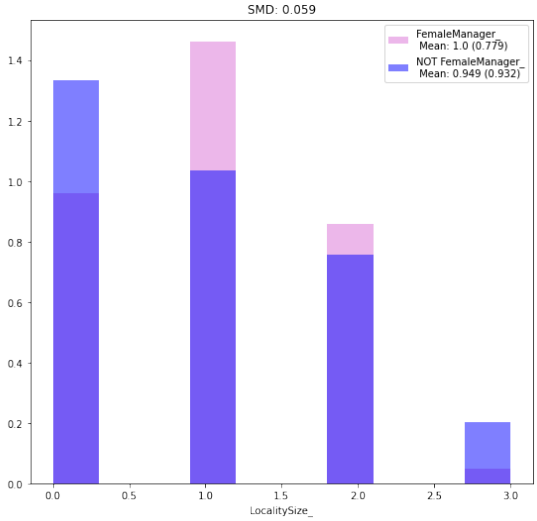
Variable	Question in BEEP
Corruption	"To what degree is/are [Corruption] an obstacle to the current operations of this establishment?
Female manager	"Is the Top Manager female?"
Locality Size	"Size of locality" City with population over 1 million / Over 250.000 to 1 million / 50.000 to 250.000 / Less than 50.000
Size	"Size" (of enterprise)
No. of Establishments	"Number of establishments that form the firm "
Subsidary	"Establishment is part of a larger firm"
Female Ownership	"What percentage of the firm is owned by females?"
Government Ownership	"What percentage of this firm is owned by each of the following: Government or State (%)"
Established	"In what year did this establishment begin operations?"
Size at Startup	"How many full-time employees did this establishment employ when it started operations? Please include all employees and managers"
Financial Risk	Over fiscal year 2012/2013, please estimate the proportion of this establishment's working capital, that is its day-to-day operations, that was financed from each of the following sources?

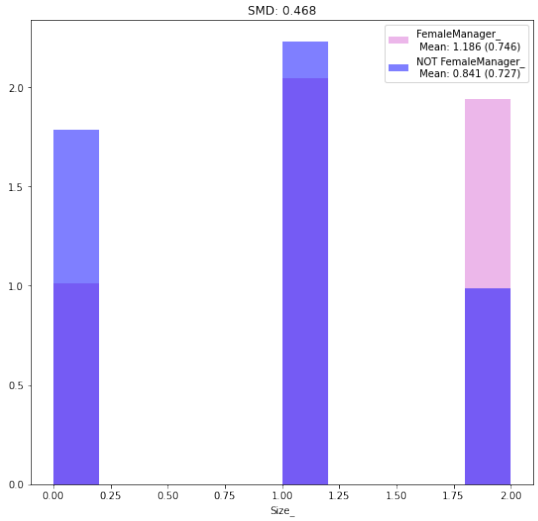
	Internal funds or retained earnings (%)
Region	“Region” - Indicator for one of India’s States
Legal Status	“What is this firm‘s current legal status?”
Industry	“Industry”

Table 2. Summary statistics of the data. There is great under-representation of the treatment group as only 7.72% of the firms have a female top manager. The statistics for the dummy variables for legal status, industry, and region are not reported here for readability. I dropped observations with NAs in any of the baseline, treatment, or outcome variables with the assumption that there is no sample selection.

	count	mean	std	min	25%	50%	75%	max
CorrPerc_	8451.0	2.123891	1.328490	0.0	1.0	2.0	3.0	4.0
FemaleManager_	8451.0	0.077151	0.266846	0.0	0.0	0.0	0.0	1.0
LocalitySize_	8451.0	0.953378	0.921513	0.0	0.0	1.0	2.0	3.0
Size_	8451.0	0.867590	0.734672	0.0	0.0	1.0	1.0	2.0
NoEstablishment	8451.0	1.588451	3.101453	1.0	1.0	1.0	1.0	200.0
Subsidary_	8451.0	0.786416	0.409861	0.0	1.0	1.0	1.0	1.0
FemaleOwner_	8451.0	3.194297	12.430898	-1.0	-1.0	-1.0	-1.0	58.0
GovOwner	8451.0	0.108153	2.058699	0.0	0.0	0.0	0.0	51.0
Established	8451.0	1995.158916	13.514619	1873.0	1989.0	1999.0	2005.0	2013.0
StartUpSize	8451.0	38.180689	87.103155	1.0	8.0	15.0	30.0	1100.0


SMD: 0.24
FemaleManager_
Mean: 2.411 (1.263)
NOT FemaleManager_
Mean: 2.1 (1.331)
CorrPerc_


SMD: 0.059
FemaleManager_
Mean: 1.0 (0.779)
NOT FemaleManager_
Mean: 0.949 (0.932)
LocalitySize_


SMD: 0.468
FemaleManager_
Mean: 1.186 (0.746)
NOT FemaleManager_
Mean: 0.841 (0.727)
Size_

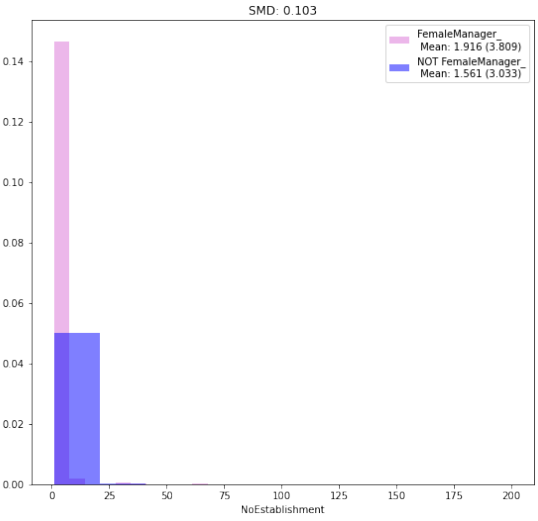

SMD: 0.103
FemaleManager_
Mean: 1.916 (3.809)
NOT FemaleManager_
Mean: 1.561 (3.033)
NoEstablishment

Figure 2. Pre-Matching balance and distribution of the variables. There is imbalance across the covariates as indicated by the absolute squared mean difference. Female ownership share in the business has the largest imbalance, followed by the firm's size, and the number of subsidiaries.

Methods

To estimate the propensity score to be matched on I estimate a logistic regression and a neural network using the covariates to predict the outcome variable. To test for overfitting, I split the full data set into training and test with a 20/80 ratio. The logistic regression takes the inputs as a linear combination that predicts an outcome variable between 0 and 1 that minimizes the mean squared error to the sigmoid function. The sigmoid function pushes the predicted values towards the extremes – zero and one. The coefficients can be interpreted as being the increase in the outcome variable with a one-unit increase in the input values.

A neural network stacks layers of nodes that are interconnected through a set of edges. Each node is thereby a function of the previous node that uses the layer's activation function as its kernel. With each edge being a parameter, the neural network allows for a high non-linearity. The parameters are trained to minimize a loss function in iterative backpropagation. The common loss function for classification class is binary cross-entropy which compares the predicted treatment probability with the actual treatment value. For each sample the binary cross-entropy calculates:

$$BCE(t,p) = -(t * \log(p) + (1-t) * \log(1-p))$$

, where t is the treatment value and p is the probability of being in the treatment group.

The number of layers and their respective nodes is arbitrary to the point that the number of parameters increases with each node with the number of nodes in the previous and following layer. As the number of parameters increases, so does the likelihood of overfitting. I use a simple network with an input layer of 65 input nodes, two middle layers with 25 and eight nodes respectively, the final layer has a single node that uses the sigmoid function that functions like the logistic regression. The network's architecture, hence, reduces the data dimensionality with each subsequent layer while allowing for the complicatedness that are shown in (Fig. 1).

There is a stark imbalance in the dataset, which may lead to the neural network predicting every firm to have a male manager. One method to avoid this is to change the weighting of each observation when calculating the BCE. Through weighting, the predicted probability of treatment is generally increased because it is "costlier" for the network to predict observations as male when they are not. Hence, there will be more false positives than false negative. The weights represent the ratio of occurrence in the entire dataset. The weights are calculated using:

$$w_g = \frac{n}{2} * (\frac{1}{n_g})$$

, where n is the total sample size and n_g is the number of samples in each treatment group g (TensorFlow, 2020). The weights are:

Weight control: 0.541406375140157

Weight treatment: 6.537717601547389

Because I am using the propensity score, I can use a one-to-one matching approach that matches with replacement within a caliper of 0.2 standard deviations as suggested by the literature (Wang, et al., 2013). Sampling with replacement usually decreases the bias (Sekhon, 2020).

Results

Upon comparison, the different propensity score models differ in accuracy, achieved balance, and final treatment effect.

A first metric for estimating how good each propensity score model is comparing the confusion matrices (Fig. 3). The model comparisons are made using a test set not used for training. As expected, the neural network is overall better at predicting which firms have a female manager as the true positive rate (bottom right in each panel) compared to the logistic regression. Both have a high false-negative rate meaning a lot of firms have been wrongly identified to have a male head. Using the weighted neural network, the true-positive rate makes a jump at cost of the false-positive rate which skyrockets, so now more female-led firms have been identified as such, while a lot of male-led firms have been mispredicted (Fig. 3, Panel C). When plotting the densities of the predicted propensity scores, the confusion matrices translate to a large overlap on the lower end of the propensity score scale for the unweighted (Fig. 4.). The weighted neural network's propensity scores look more like what is expected of a "good" model capturing bias in the treatment assignment as the probability of being in the treated group is low for the control firms and high for the treatment firms. However, this may be misleading if in reality the predictors of having a female manager are not the variables chosen here.

When interpreting the models, the coefficients for the logistic regression and shap values can be of some indication of covariate's impact on the model output. Shap values are calculated using the mean of the marginal contribution adding the respective covariate has when included in the prediction across the input values (Lundberg & Lee, 2017). As shown in Fig. 5, the largest impact was caused by having a female owner. This makes sense, as often times the owner of the firm is also the manager, so if there is a female owner, then the manager is likely female too. The

logistic regression also shows that female ownership is a significant predictor of the propensity score (See Appendix for regression table). However, the standard errors are extremely high for the indicator variables, making them nonsignificant.

After matching all propensity score models are able to improve the balance on the covariates (Fig. 6). As shown by the mean SMD, the logistic regression performs best on reducing average SMD. However, using the average is to be taken with caution, as the relative weight of each covariate in biasing the effect is unknown, especially when including the fact that the industry, legal status, and region indicators are also balanced on (See Appendix A). The unweighted neural network has a balancing effect on the majority of the covariates but not or the firm's year of establishment and the percentage of government ownership. The unweighted models each drop five treatment observations, while the weighted neural network drops none. Neither drops a large portion of the treated sample, so the final sample sizes are roughly 1300.

As a final step, I estimate an Ordinary Least Squares model with each matched dataset (Tab. 3). that shows that all techniques agree that the null hypothesis that female management does not affect the perception of corruption cannot be rejected. The coefficients are insignificant at the conventional thresholds, and from the non-indicator variables, only the financial risk seems to contribute robustly to the perception of corruption. The more risk-taking the less the firm sees corruption to be a problem. However, none of these results are valid and by far not causal if the assumptions of the propensity score are estimation are not fulfilled.

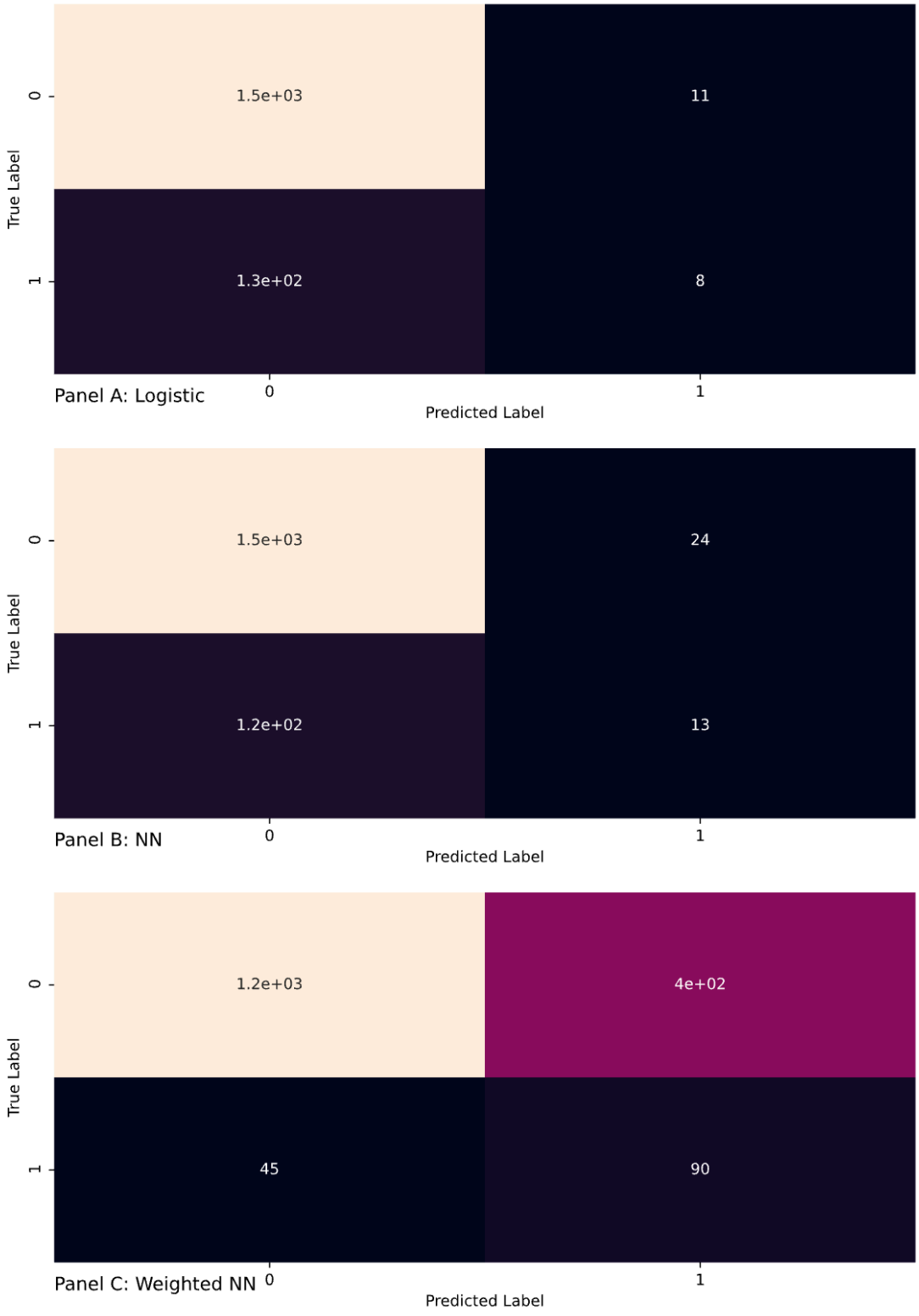


Figure 3. Confusion matrices of the propensity score models. The unweighted logistic regression (Panel A) and neural network (Panel B) have a high false negative rate while the weighted neural network has a high false positive rate (Panel C).

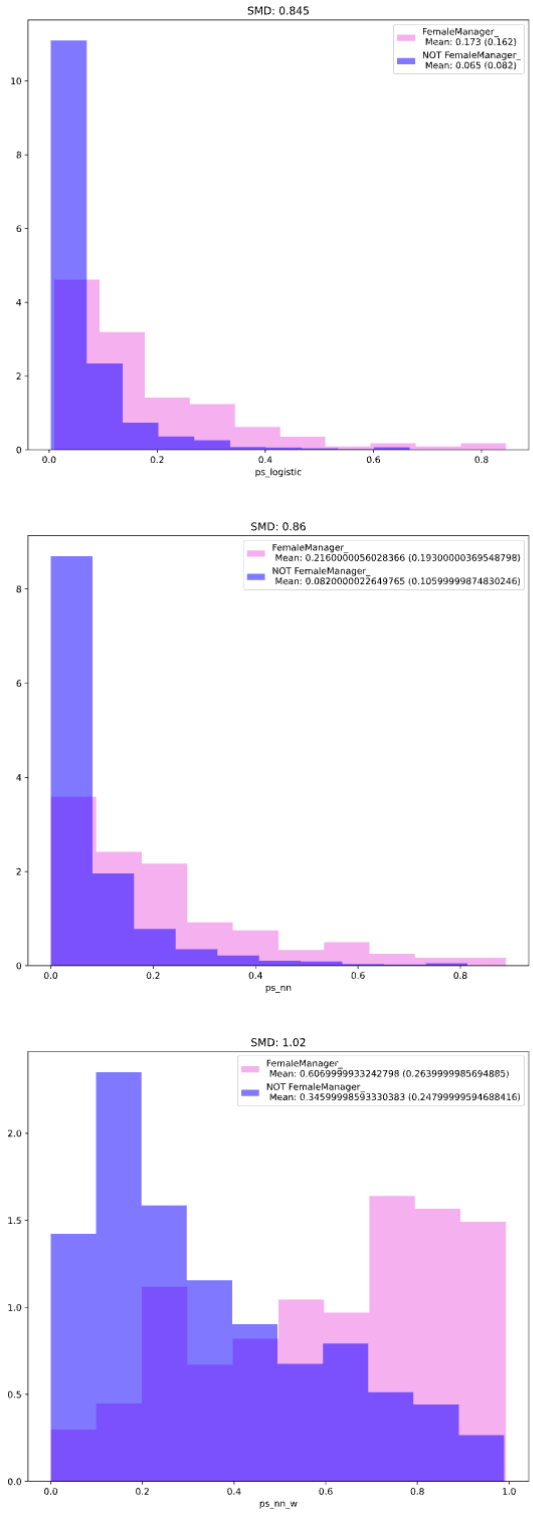
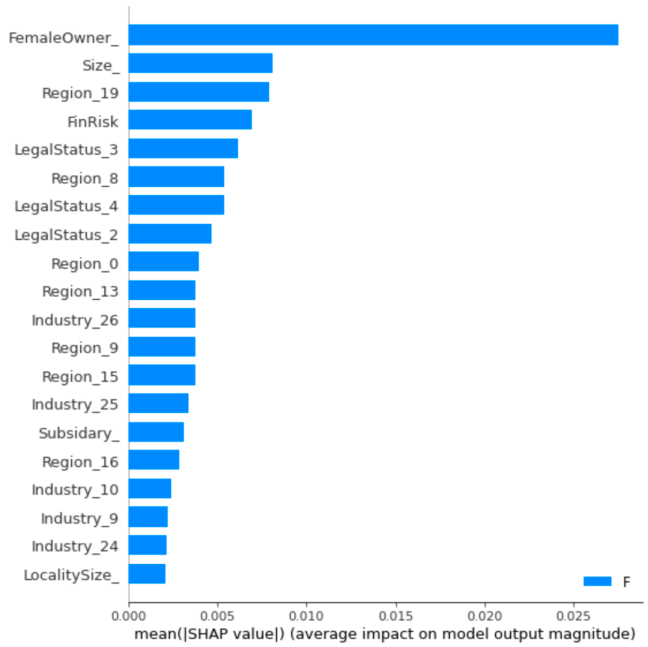
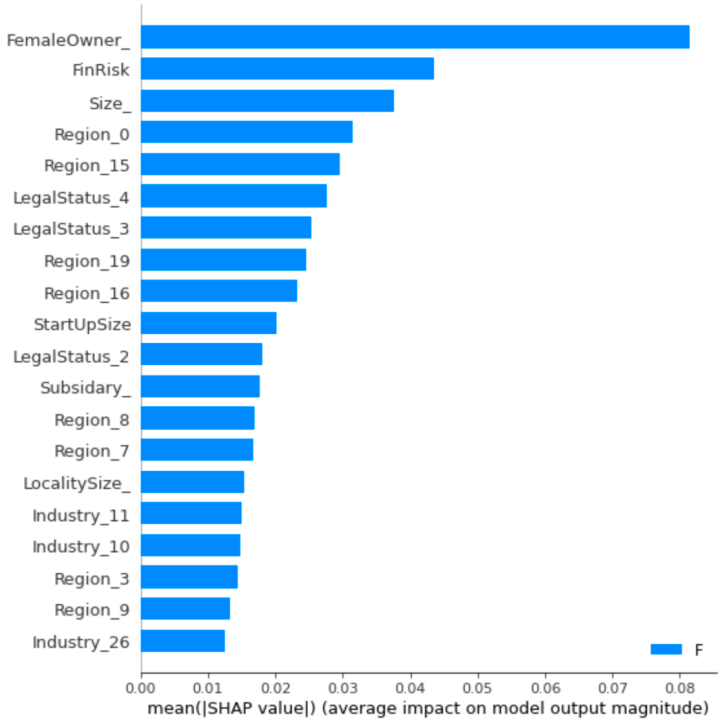


Figure 4. Propensity score distributions. The logistic regression (top) is slightly worse at discriminating between the propensity scores of treatment and control group than the neural net (middle). The weighted neural net (bottom) distinguishes well between the classes.



Panel A: Unweighted NN



Panel B: Weighted NN

Figure 5. Shap Values for the two neural networks. In both the amount of female ownership in the firm has the largest effect. Size and Financial risk are also ranked highly in both. Panel A shows the unweighted neural network and Panel B the weighted neural network. Note that the x-axis for the two panels are not the same.

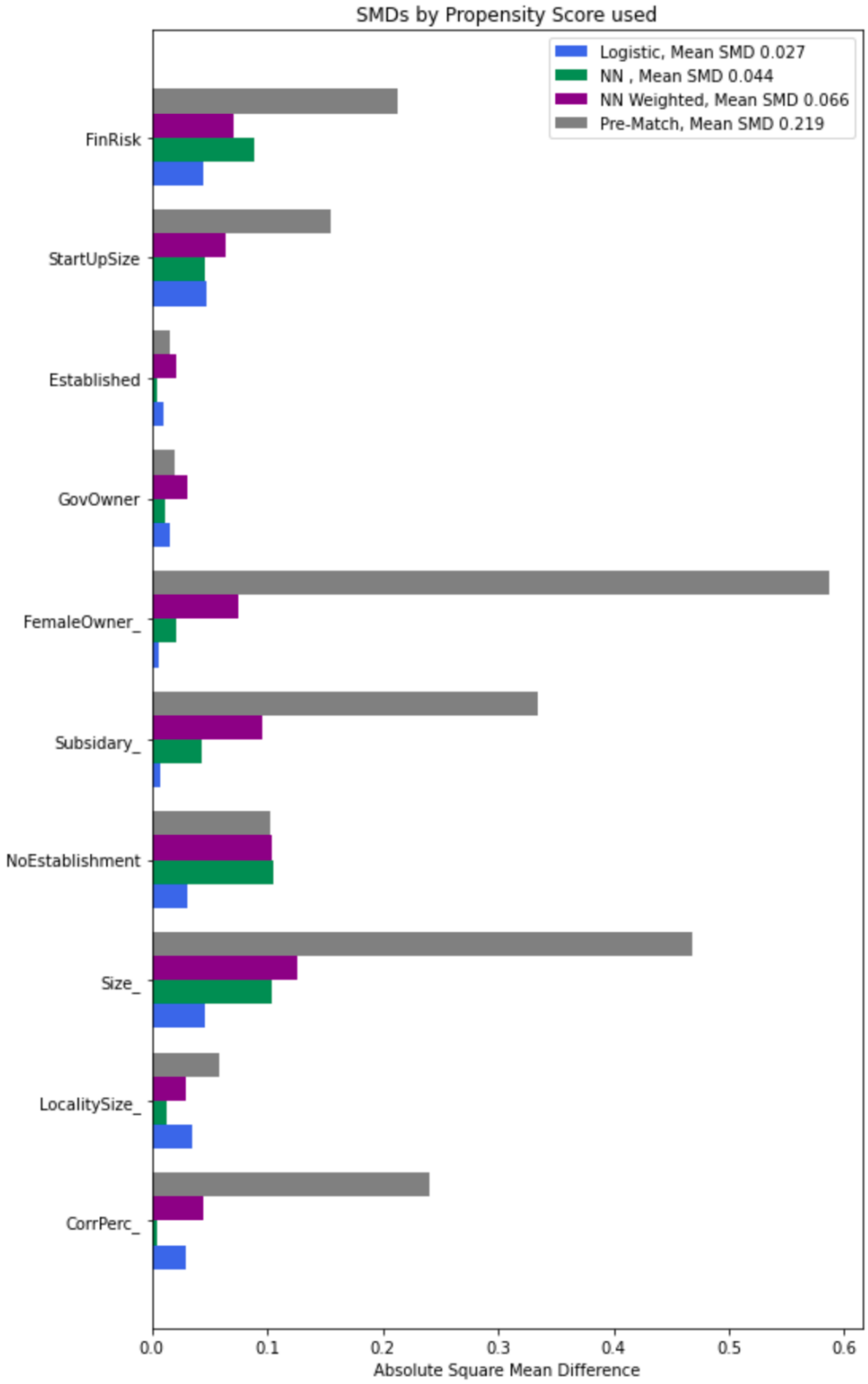


Figure 6. Standard Mean Difference by baseline covariate. The balance generally improves when matching on the propensity score. SMDs of the dummy baseline covariates are given in the appendix.

Table 3. Final OLS results. Female Management does not significantly affect the perception of corruption. For easier readability, the fixed effects for the industries, regions, and legal status are given in the appendix. Standard errors are shown below each coefficient.

Dependent variable: CorrPerc			
	Logistic PS	NN PS	Weighted NN PS
Established	0.003	0.008***	0.004*
	-0.003	-0.003	-0.003
FemaleManager_	0.032	0.035	-0.028
	-0.065	-0.064	-0.064
FemaleOwner_	0.002	0.003	0.002
	-0.002	-0.002	-0.002
FinRisk	-0.004***	-0.004***	-0.005***
	-0.001	-0.001	-0.001
GovOwner	0.040**	0.035**	0.048***
	-0.016	-0.016	-0.013
LocalitySize_	-0.014	0.056	0.048
	-0.051	-0.051	-0.051
NoEstablishment	0.002	-0.004	-0.012
	-0.005	-0.012	-0.011
Size_	-0.014	0.072	0.011
	-0.054	-0.054	-0.052
StartUpSize	-0.001	-0.001	0
	0	0	0
Subsidary_	-0.03	-0.011	-0.044
	-0.076	-0.078	-0.078
Observations	1,272	1,294	1,304
R^2	0.282	0.291	0.287
Adjusted R^2	0.245	0.255	0.252
Residual Std. Error	1.141 (df=1209)	1.134 (df=1232)	1.126 (df=1242)
F Statistic	7.656*** (df=62; 1209)	8.271*** (df=61; 1232)	8.195*** (df=61; 1242)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Discussion

Even though the results are in "agreement" that the OLS model fails to reject the null hypothesis, the propensity score models are not perfect. Both the unweighted and weighted models are insufficient in accurately gender of the firm's manager. While the weighted neural network was able to increase the number of firms correctly identified to have female leadership, this came at the cost of increasing the false positive rate. As shown in Fig. 6, this did not actually improve the balance of the covariates compared to the unweighted methods. This is because the approximation of propensity scores is even more distorted when including weights and potential covariates', potentially inducing more bias.

Another problem with the model is induced through the large number of dummy variables. While they are certainly important predictors of the propensity score, the industry, region, and legal dummies lead to the data being blocked into smaller and smaller chunks meaning certain combinations of variables occur with small n. Hence, only the combination of the dummy variables is enough for the model to estimate the outcome accurately. Hence, it overfits and does not disentangle the trend of the treatment variable. When re-running the final stage OLS without the dummy variables, the F-statistics turn insignificant, meaning the model is not any better as an intercept model. Hence, the final OLS models' R^2 are driven by the dummy variables rather than the predictive variables. This illustrates a problem with one-to-one matching, as it drastically reduces the sample size. The problem of small n is also an issue, as shown in by the large coefficient standard errors (see Appendix). When calculating standard errors, the assumption is that the sample is able to cover the true population appropriately. With a small number of samples, this does not apply anymore, and the standard errors increase. Hence, in a further iteration of the model, finding accurate groupings for the regions and industries that would reduce the number of

combinations is a critical step in improving the model. Simply leaving them out of either the model or the propensity score estimation is not an option either, as it would mean omitting a relevant confounder.[1]

As noted before, the propensity scores' validity depends on the model that is specified to estimate it. I used the most exogenously defined variables that are not causal children of either corruption or female management to build the model. However, if the set of variables is not an appropriate set of variables to predict the variables that lead to female management, then the propensity score is invalid (Pearl, 2009). No model (whether logistic or neural network) can achieve an accurate representation of the propensity score without receiving the proper covariates. The literature provided above gives an idea of which baseline covariates should be included. I choose a relatively generic set, but there may be additional variables that may influence treatment and outcome. One such example might be the level of conservative values, the owner or founder of the business holds, or the number of females in the firm previous to the appointment of the current manager. Both are not in the dataset. Similarly, including variables that are not parents of the treatment should be excluded. Examples of this could be the financial risk that the firm takes. I am assuming the risk is taken by the owner (as this is how I measure risk), but it could as well be the manager's decision. Including or excluding variables from the set of propensity score predictors increases bias (Caliendo & Kopeinig, 2008) (Pearl, 2009) (King & Nielsen, 2019). It is hence the largest threat to the causal and even the descriptive validity of the results.

---

[1] #sampling: Throughout the assignment I try to control for the small number of samples in the treatment group through, e.g., weighting. I implement, compare, and evaluate propensity score matching techniques which allows to create a balanced sample. By understanding the structure of my data and the sampling that takes place, I am able to understand the limitations of the matched data in estimating robust regression coefficients.

Usually, the only way the baseline covariates are chosen is through background research and common sense. However, in an extension to this study, I could use causal discovery using the full dataset. While this might be computationally expensive, it would give me a set of baseline covariates that could be relevant that are in the survey. Relying solely on a causal discovery algorithm, such as the PC or FCI algorithms, would not be sufficient, and research has to back-up each variable. Still, it could be an additional source of information.

The method's motivation was to test out different propensity score estimation models using real-world data that is not “perfect” or extensively studied like the Lalonde dataset. A useful extension would be how the estimated propensity scores fair with different matching algorithms such as matching without replacement, with one-to-many matching (which might help with the problem of too few samples in each subgroup), matching on the propensity score and the covariates, or matching on only the covariates using the Mahalanobis distance.

The final model is evaluated using a simple OLS. However, the outcome variable is a discrete variable, even though OLS is discrete. If it was the core purpose to find the probability of for each level, I could have used a multinomial model, but the variable is ordered so to just see if there is an effect, the simpler OLS suffices.

I dropped observations with NAs in any of the baseline, treatment, or outcome variables with the assumption that there is no sample selection. Testing for bias in the missingness could be tested with a sample selection adjusting model such as Heckman or seeing if there is any significant predictor for missingness when using a logistic regression. Imputing missing values could also help with the small sample size.

Another way to find the causal effect without relying on observing all covariates would be to find a credible instrument for having a female manager. A possible way to achieve this would

be using an encouragement design of schooling business owners on the benefits of having female managers and using their random assignment to the encouragement group as the instrument for having female managers. A non-experimental approach could be to use a change in policy that allows for easier access to microfinance for women in one state versus another to be the encouragement and hence instrument. Still finding an instrument for such an endogenous variable such as gender in the business context will be difficult to defend against exclusion restriction (no effect of instrument on outcome), relevance (instrument has predictive power of treatment), and independence (no confounder between instrument, treatment/outcome)[2].

[2] #observationalstudy: I go through the design process of building up an observational study from research question, to finding data, literature review, methodology, results, and thorough analysis of potential improvements to the method. I also provide extensions such as using an instrumental variable from a policy intervention as a non-trivial alternative.

Conclusion

Overall, the comparison of the different propensity scoring methods did not lead to conclusive results about the effect of female management on the perception of corruption to be an obstacle for the firm. Because there are not enough specific baseline covariates, the inclusion of dummy variables for regional, industry, and legal status differences could not be avoided to account for confoundedness. Through a thorough analysis of the results, I find that the data becomes blocked into too many subgroups. The blocking would have gone unnoticed without proper inspection of the propensity score and final model as the matching seems to have brought balance to the dataset. This leads to high standard errors and overfitting of the models. Hence, the results have to be taken with caution. Further analysis should include grouped regions and industry indicators, could make use of causal discovery to find relevant covariates, or use a strong instrument to strengthen the causal claim of the paper.

Works cited

Abdixhiku, L., Krasniqi, B., Pugh, G., & Hashi, I. (2017). Firm-level determinants of tax evasion in transition economies. *Economic Systems*, 354-366.

Alatas, V., Cameron, L., Chaudhuri, A., Erkal, N., & Gangadharan, L. (2009). Gender, Culture, and Corruption: Insights from an Experimental Analysis. *Southern Economic Journal*, 663-680.

Breen, M., Gillanders, R., McNulty, G., & Suzuki, A. (2017). Gender and Corruption in Business. *The journal of development studies*, 1486-1501.

Cabral, A. C., Kotsogiannis, C., & Myles, G. (2019). Self-Employment Income Gap in Great Britain: How Much and Who? *CESifo Economic Studies*, 84–107.

Caliendo, M., & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of economic surveys*, 31-72.

Dollar, D., Fisman, R., & Gatti, R. (2001). Are women really the "fairer" sex? *Journal of Economic Behavior & Organization*, 423-429.

Frank, B., Lambsdorff, J. G., & Boehm, F. (2011). Gender and corruption: Lessons from laboratory corruption experiments. *The European Journal of Development Research*, 59-71.

Hurst, E., Li, G., & Pugsley, B. (2014). ARE HOUSEHOLD SURVEYS LIKE TAX FORMS? EVIDENCE FROM INCOME UNDERREPORTING OF THE SELF-EMPLOYED. *The Review of Economics and Statistics*, 19-33.

King, G., & Nielsen, R. A. (2019). Why propensity scores should not be used for matching. *Political Analysis*.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* , 4765-4774.

Pearl, J. (2009). Understanding Propensity Scores. In J. Pearl, *Causality* (pp. 348-352). Cambridge University Press.

Pissarides, C. A., & Weber, G. (1989). AN EXPENDITURE-BASED ESTIMATE OF BRITAIN'S BLACK ECONOMY. *Journal of Public Economics*, 17-32.

Sekhon, J. S. (2020, Feb 06). *Multivariate and propensity score matching with balance optimization.* From Package 'Matching'.

Slemrod, J., & Yitzhaki, S. (2002). Tax avoidance, evasion, and administration. *Handbook of public economics*, 1423-1470.

Svensson, J. (2005). Eight questions about corruption. *Journal of Economic Perspectives*, 19-42.

TensorFlow. (2020, Nov 21). *Classification on imbalanced data*. From TensorFlow Core: https://www.tensorflow.org/tutorials/structured_data/imbalanced_data

Wang, Y., Cai, H., Li, C., Jiang, Z. W., Song, J., & Xia, J. (2013). Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study. *PloS one*, e81045.

World Bank. (2014). *India Enterprise Survey (ES)*. Retrieved Nov 25, 2020 from Ref. IND_2014_ES_v01_M: [.DTA file]

Appendix A

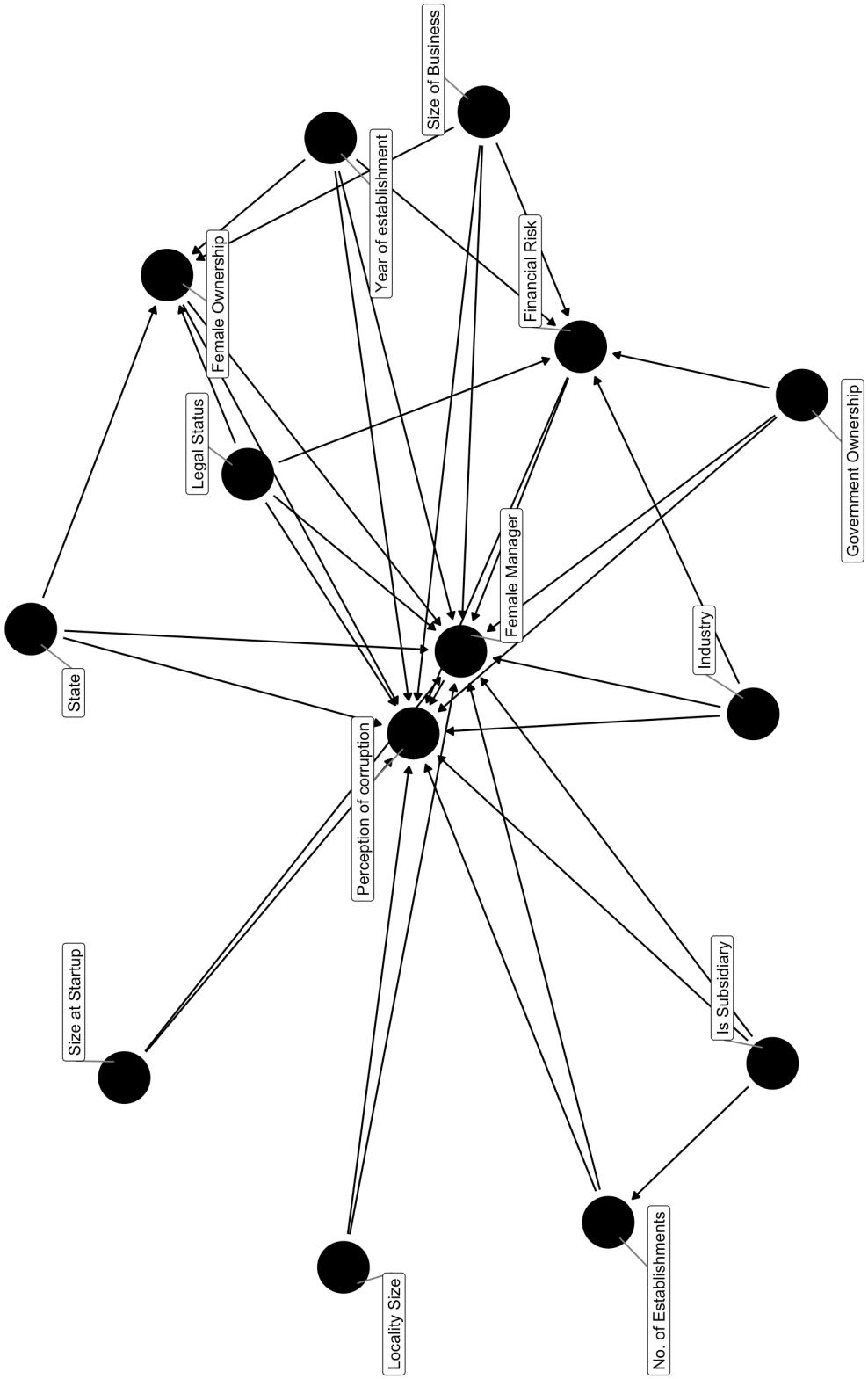


Figure 7. DAG but in larger for better readability.

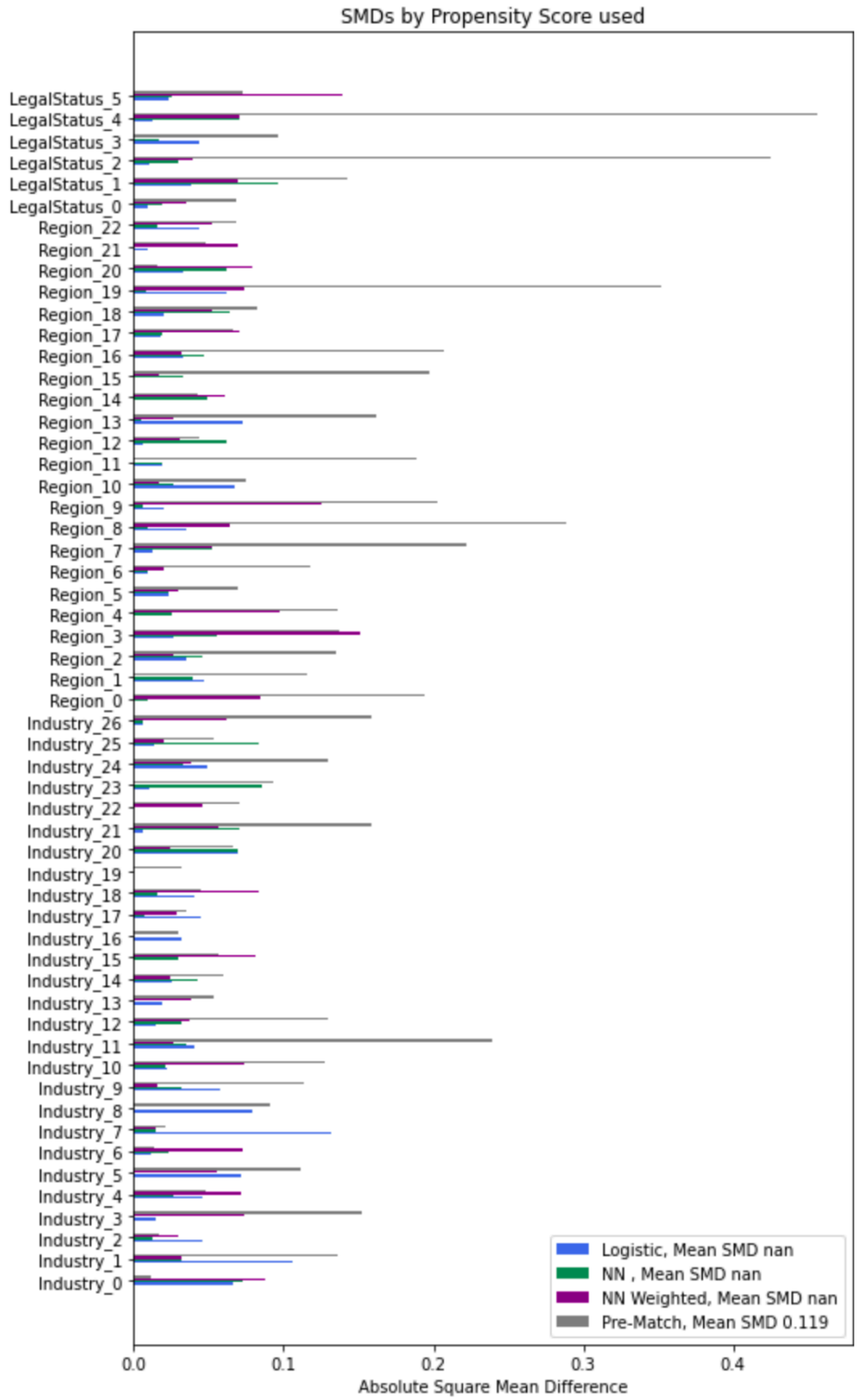


Figure 8. Balance across the dummy variables. Generally the propensity score matchings improve balance. As shown by the "nan" values in the estimation of the mean SMD, there are some categories such as Industry 19 where there are no observations left after matching. As addressed in the discussion section the small n in the large number of these blocks causes issues of overfitting.

Table 4. Coefficients of the logistic regression propensity model. The standard errors on the industry, region, legal status dummies are likely so high because there are very few observations in the individual combinations. This allows the coefficients to vary widely across the sample and hence the standard errors skyrocket. The reported coefficients are calculated using the stats model package making them marginally different to the ones calculated by sklearn which I used for the further analysis. The difference is ignorable.

Dependent variable: FemaleManager_	
	(1)
LocalitySize_	-0.098
	(0.078)
Size_	0.390***
	(0.084)
NoEstablishment	0.003
	(0.014)
Subsidary_	-0.225*
	(0.120)
FemaleOwner_	0.043***
	(0.003)
GovOwner	-0.012
	(0.024)
Established	-0.000
	(0.004)
StartUpSize	-0.001
	(0.001)
FinRisk	-0.003
	(0.002)
Industry_0	0.999
	(29072131.599)
Industry_1	-1.018

	(29910545.763)
Industry_2	0.502
	(29425963.625)
Industry_3	0.991
	(31405030.073)
Industry_4	0.411
	(29704733.595)
Industry_5	-0.301
	(30246210.955)
Industry_6	0.847
	(29266708.001)
Industry_7	0.554
	(30284274.480)
Industry_8	-13.508
	(29914301.971)
Industry_9	0.943
	(29633391.225)
Industry_10	0.287
	(30109965.649)
Industry_11	-0.457
	(30582790.308)
Industry_12	0.596
	(31613989.283)
Industry_13	0.653
	(29568845.243)

Industry_14	0.527
	(31048716.242)
Industry_15	0.609
	(29945278.656)
Industry_16	0.345
	(29684862.922)
Industry_17	0.412
	(29828389.133)
Industry_18	1.116
	(30196314.960)
Industry_19	-11.297
	(29162197.746)
Industry_20	1.294
	(30372286.875)
Industry_21	1.325
	(29590691.038)
Industry_22	0.807
	(30169317.800)
Industry_23	0.782
	(29295812.676)
Industry_24	1.323
	(29753250.485)
Industry_25	1.394
	(29833097.909)
Industry_26	1.709

	(29963724.042)
Region_0	-2.030
	(45088086.702)
Region_1	-1.790
	(44988685.186)
Region_2	-1.729
	(43803077.805)
Region_3	-1.553
	(43577992.471)
Region_4	-1.501
	(44527908.156)
Region_5	-0.850
	(43398376.104)
Region_6	0.333
	(44153618.240)
Region_7	-1.668
	(47280679.816)
Region_8	0.217
	(45321379.419)
Region_9	0.413
	(44653093.800)
Region_10	-0.045
	(46314830.498)
Region_11	-1.579
	(43803077.805)

Region_12	-0.897
	(43813340.479)
Region_13	-0.019
	(44587602.384)
Region_14	-0.880
	(43846674.807)
Region_15	-1.444
	(45449323.877)
Region_16	-2.426
	(41696350.062)
Region_17	-0.984
	(45295333.202)
Region_18	-1.582
	(45045691.203)
Region_19	0.201
	(47226371.687)
Region_20	-0.584
	(43727749.543)
Region_21	-1.010
	(45277135.351)
Region_22	-1.580
	(47306422.915)
LegalStatus_0	-2.155
	(12842834.204)
LegalStatus_1	-2.681

	(12842834.204)
LegalStatus_2	-2.851
	(12842834.204)
LegalStatus_3	-3.115
	(12842834.204)
LegalStatus_4	-2.194
	(12842834.204)
LegalStatus_5	-3.169
	(12842834.204)
Observations	6,760
R^2	
Adjusted R^2	
Residual Std. Error	1.000 (df=6697)
F Statistic	(df=62; 6697)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 5. Coefficients of final model dummy variables. As shown, some of them are highly significant, which is how they collectively lead to a significant F-statistic in the model. Standard errors are shown below each coefficient.

	Logistic PS	NN PS	Weighted NN PS
Industry_0	-0.125	-1.665*	-0.765
	-0.845	-0.85	-0.832
Industry_1	-0.008	-1.582	-0.496
	-0.898	-1.029	-0.998
Industry_10	-0.452	-1.955**	-1.018
	-0.852	-0.853	-0.843
Industry_11	-0.319	-2.041**	-0.774
	-0.881	-0.878	-0.866
Industry_12	-0.444	-1.962**	-0.927
	-0.859	-0.869	-0.848
Industry_13	-0.263	-2.013**	-0.848
	-0.848	-0.854	-0.837
Industry_14	-0.337	-1.985**	-1.08
	-0.836	-0.843	-0.822
Industry_15	-0.203	-1.920**	-0.779
	-0.844	-0.853	-0.836
Industry_16	-1.796*	-2.788**	-1.079
	-1.063	-1.129	-1.146
Industry_17	-0.427	-1.906**	-0.889
	-0.856	-0.857	-0.835
Industry_18	-0.074	-2.082**	-1.173

	-0.889	-0.903	-0.892
Industry_19	0	-0.000**	0
	0	0	0
Industry_2	-0.231	-2.034**	-1.051
	-0.849	-0.856	-0.833
Industry_20	-0.523	-2.361***	-1.232
	-0.85	-0.866	-0.848
Industry_21	-0.861	-2.728***	-1.565*
	-0.867	-0.873	-0.857
Industry_22	-0.889	-2.300**	-0.526
	-0.91	-0.9	-0.91
Industry_23	-0.156	-2.161**	-1.237
	-0.893	-0.887	-0.878
Industry_24	-0.566	-2.396***	-1.033
	-0.854	-0.861	-0.835
Industry_25	-0.372	-2.013**	-1.17
	-0.863	-0.862	-0.846
Industry_26	-0.64	-2.482***	-1.301
	-0.872	-0.872	-0.856
Industry_3	-0.123	-1.914**	-0.776
	-0.858	-0.867	-0.845
Industry_4	-0.165	-2.039**	-0.926
	-0.874	-0.895	-0.88

Industry_5	-0.565	-2.385**	-1.101
	-0.908	-0.945	-0.909
Industry_6	-0.516	-2.534***	-1.217
	-0.88	-0.876	-0.857
Industry_7	0.049	-1.929**	-0.552
	-0.905	-0.883	-0.859
Industry_8	-1.129	0.000**	0
	-1.12	0	0
Industry_9	-0.426	-2.162**	-0.783
	-0.848	-0.851	-0.839
LegalStatus_0	-2.467	-9.177***	-4.625
	-3.615	-3.507	-3.442
LegalStatus_1	-1.948	-9.073**	-3.988
	-3.634	-3.52	-3.448
LegalStatus_2	-1.892	-8.804**	-3.89
	-3.634	-3.52	-3.451
LegalStatus_3	-1.796	-8.742**	-3.901
	-3.632	-3.513	-3.45
LegalStatus_4	-1.873	-8.826**	-3.926
	-3.629	-3.517	-3.448
LegalStatus_5	-1.588	-8.716**	-3.967
	-3.662	-3.527	-3.486
Region_0	-1.504	-3.153***	-2.173**

	-0.982	-0.948	-0.934
Region_1	-0.14	-2.438**	-1.276
	-1.004	-0.978	-0.97
Region_10	0.091	-1.838**	-0.86
	-0.968	-0.934	-0.915
Region_11	-0.637	-2.126**	-0.965
	-1.021	-0.994	-0.979
Region_12	-1.406	-3.273***	-2.106**
	-0.961	-0.935	-0.915
Region_13	-0.538	-2.396***	-1.231
	-0.94	-0.912	-0.893
Region_14	0.041	-1.921**	-0.569
	-0.957	-0.926	-0.914
Region_15	-0.832	-2.655***	-1.084
	-0.959	-0.93	-0.914
Region_16	-1.55	-3.543***	-2.185**
	-1.014	-0.975	-0.957
Region_17	0.366	-1.648*	-0.123
	-0.959	-0.934	-0.925
Region_18	-0.463	-2.281**	-0.78
	-0.968	-0.953	-0.93
Region_19	-0.368	-2.186**	-0.852
	-0.953	-0.922	-0.909

Region_2	-0.397	-2.319**	-1.097
	-0.967	-0.942	-0.936
Region_20	0.502	-1.307	-0.007
	-0.957	-0.927	-0.913
Region_21	0.322	-1.614*	-0.326
	-0.99	-0.958	-0.943
Region_22	-1.650*	-3.475***	-1.956**
	-0.948	-0.915	-0.906
Region_3	-0.937	-2.555***	-1.109
	-0.982	-0.956	-0.931
Region_4	-1.134	-2.491***	-1.492
	-0.989	-0.957	-0.933
Region_5	-0.155	-1.777*	-0.6
	-0.944	-0.919	-0.9
Region_6	-0.174	-1.860**	-0.683
	-0.96	-0.93	-0.914
Region_7	-1.391	-3.065***	-1.758*
	-0.984	-0.963	-0.951
Region_8	0.293	-1.709*	-0.426
	-0.955	-0.92	-0.911
Region_9	0.099	-1.706*	-0.641
	-0.976	-0.945	-0.92