Blöthner, Simon; Larch, Mario

Working Paper

Economic Determinants of Regional Trade Agreements Revisited Using Machine Learning

CESifo Working Paper, No. 9233

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Blöthner, Simon; Larch, Mario (2021) : Economic Determinants of Regional Trade Agreements Revisited Using Machine Learning, CESifo Working Paper, No. 9233, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/245414

CESifo WORKING PAPERS

9233
2021

August 2021

Economic Determinants of Regional Trade Agreements Revisited Using Machine Learning

Simon Blöthner, Mario Larch

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9233*

Economic Determinants of Regional Trade Agreements Revisited Using Machine Learning

Abstract

While traditional empirical models using determinants like size and trade costs are able to predict RTA formation reasonably well, we demonstrate that allowing for machine detected non-linear patterns helps to improve the predictive power of RTA formation substantially. We employ machine learning methods and find that the fitted tree-based methods and neural networks deliver sharper and more accurate predictions than the probit model. For the majority of models the allowance of fixed effects increases the predictive performance considerably. We apply our models to predict the likelihood of RTA formation of the EU and the United States with their trading partners, respectively.

JEL-Codes: F140, F150, C450, C530.

Keywords: Regional Trade Agreements, neural networks, tree-based methods, high-dimensional fixed effects.

Simon Blöthner
University of Bayreuth / Germany
Simon.Bloethner@uni-bayreuth.de

Mario Larch
Department of Law and Economics
University of Bayreuth / Germany
mario.larch@uni-bayreuth.de

July 30, 2021
Acknowledgements: To be added.

1 Introduction

Regional Trade Agreements (RTAs), capturing Customs Unions (CUs), Free Trade Agreements (FTAs), Economic Integration Agreements (EIAs), Partial Scope Agreements (PSA), as well as combined agreements (CU&EIAs, FTA&EIAs, and PSA&EIAs) are the most prominent form of closer economic cooperation of countries. The number of trade relationships with an RTA, starting from 1960, has increased up to 4,721 by 2019 as shown in Figure 1. While early agreements mainly dealt with reducing tariffs, nowadays trade agreements cover also issues related to health, the environment, and labor markets, to name just a few. Additionally, EIAs covering services are increasingly popular. Starting from 1960, the number of trade relationships with an EIA has increased from 0 to 493. One plausible reason for the surge of RTAs is the slowdown of progress of multilateral liberalizations under the umbrella of the World Trade Organization (WTO). Given the current situation of the WTO, forming RTAs may even become more attractive. However, the use of RTAs differs widely across countries. Some countries have over 90 RTAs, like members of the European Union (105), Tunisia (105) and Chile (91), while other countries have no or only a few RTAs that are notified to the WTO as illustrated in Figure 2. What explains this heterogeneity? And if a country considers negotiating a new RTA, which trading partners should it focus on? Understanding the determinants of RTAs is important in order to understand RTA formation and to be able to predict potential new RTAs.

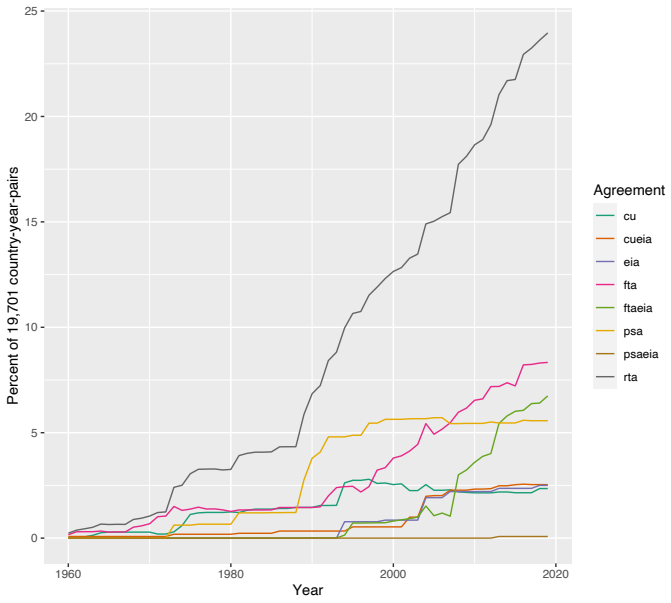

Figure 1: Development of the number of trade relationships with an RTA over the years 1960 - 2019.

The seminal article that took up this question is Baier and Bergstrand (2004), who motivate seven hypothesis based on a simulated multi-country, two-sector trade theory model with Dixit–Stiglitz preferences and increasing returns to scale. They test the seven hypothesis estimating a probit model using a cross-section of 54 countries in the year 1996 and are able to explain about 70 % of the variation in RTA conclusion. Egger and Larch (2008) reproduce this specification for 146 countries and the year 2005, explaining a bit less than 30 % of the variation in RTA

formation. The drop in the explanatory power of the probit model may be explained by the increase in the number of countries and the substantial increase of RTA formation in the last two decades, which both add substantial heterogeneity and makes it harder for a comparably parsimonious probit model to explain RTA formation.[1]

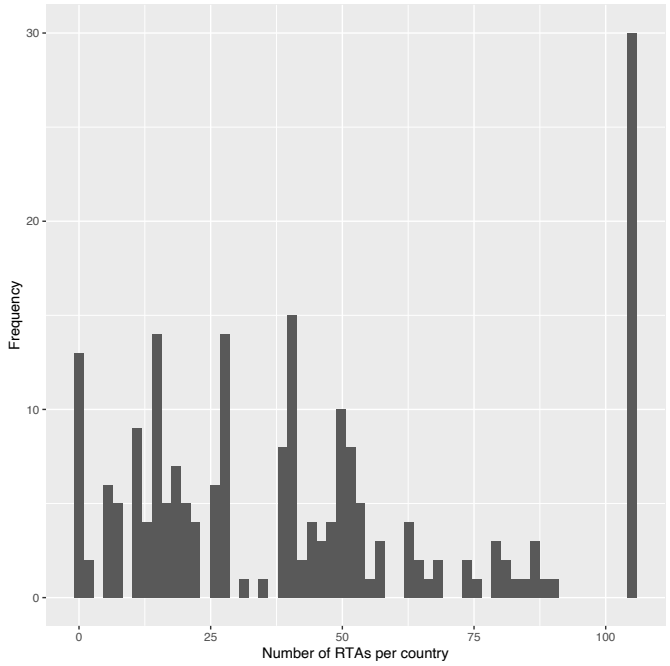

Figure 2: Frequency plot of the number of RTAs per country in 2018.

While Baier and Bergstrand (2004) motivated the determinants of RTAs aptly with their model for the cross-section of 1996, the empirical specifications of Baier and Bergstrand (2004) and Egger and Larch (2008) relied on a comparable simple probit specification based on the standard gravity variables including proxies for trade costs and size of countries. While these measures are well motivated and clearly relevant for explaining RTA determinants, they are combining country specific information, like the GDP of countries, via strong functional form assumptions that are vulnerable to miss-specification. The literature explaining the determinantes of RTAs followed this approach. However, such specifications are not capable of capturing more substantial non-linearities and complex interactions between explanatory variables, which were already visible from the simulated model in Baier and Bergstrand (2004). Figures 3 and 4 display the share of RTAs against the distance and the difference in GDP of trading pairs, respectively. As can be seen from these plots, the relationship between RTA formation and the determinants is not a simple linear relationship. However, typical empirical models trying to explain RTA determinants do not allow for such non-linearities. Given the complex nature of the world around us and the economic relationships governed by it, it is highly unlikely that traditional, linear approaches are the most adept at modeling these relationships. One of their drawbacks is that the researcher has to specify a functional form which puts strong restrictions

[1]Since Baier and Bergstrand (2004) many papers refined the specification, including for example neighboring effects like Egger and Larch (2008), Chen and Joshi (2010), Baldwin and Jaimovich (2012), and Baier et al. (2014) or political economy motives like Facchini et al. (2013), Maggi and Rodríguez-Clare (2007), Liu (2008), and Liu and Ornelas (2014). See Maggi (2014) for an excellent survey.

on what can be studied. This backs the case for some form of epistemic humility in our perceived understanding of rules guiding our surroundings.

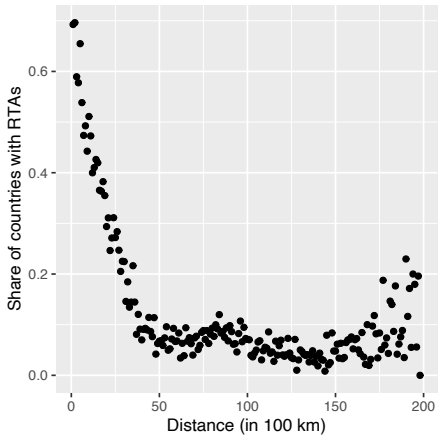

Figure 3: Non-linear relationship between the distance and the share of countries that have an RTA at a particular distance.

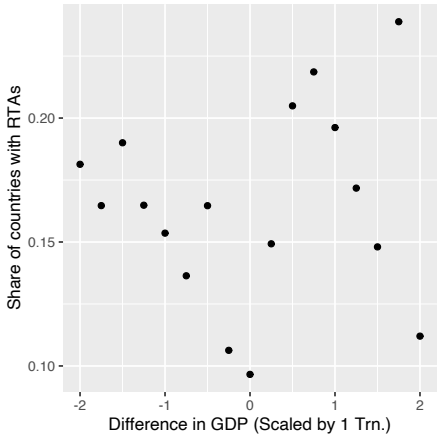

Figure 4: Non-linear relationship between the absolute difference of countries' GDP and the share of countries that have an RTA in that difference bin.

Motivated by these facts, we investigate, using developments in machine learning, whether non-linearities in explanatory variables and machine detected patterns from the data help to improve the fit and predictive power of RTA formation. We find that the probit model allowing only for limited non-linearities in the explanatory variables delivers predictions that are not very sharp for many relationship. On the other hand, the fitted tree-based methods and neural networks (NNs) deliver sharp predictions and display the considerable non-linearities and complex interactions between explanatory variables at play. For both, the probit (estimated using recent developments in the estimation of generalized linear models with multiple fixed effects) and the NN, the allowance of exporter and importer fixed effects increases the out-of-sample predictive performance considerably. Our analysis shows that our best performing machine learning algorithm (the gradient boosting machine) outperforms the probit model in true negatives/positives by 17.9/18.6 percentage points when not taking into account fixed effects and by 0.9/41.6 percentage points when taking into account fixed effects (the NN). Employing these insights, we then probe existing theories around RTA formation using the fitted machine learning models. It is of course not possible to interpret coefficients as performed in Baier and Bergstrand (2004) or Egger and Larch (2008), but we are instead trying to detect patterns arising from the fitted models, which we have shown to be a better representation of our data and the underlying data generating process than conventional linear models used in the literature, and to see if they are in line with economic theory.

In a second step we turn to predicting the likelihood of RTA formation with trading partners of actual ongoing negotiations for the European Union (EU) as well as to predict the current existing agreements of the United States.

While there are already a lot of good textbooks about machine learning, see for example Hastie et al. (2009), Murphy (2012), James et al. (2013), Efron and Hastie (2016), Goodfellow et al. (2016), and Taddy (2019), these methods are only slowly applied in the field of economics.

A good summary is given by Varian (2014). Three articles published from a symposium on "Recent Ideas in Econometrics" in the *Journal of Economic Perspectives* also provide nice overviews. Specifically, Athey and Imbens (2017) discuss the possibility of machine learning algorithms to improve credibility of policy evaluation, Stock and Watson (2017) mention the use of machine learning methods with time series data as one of the challenges ahead, and Mullainathan and Spiess (2017) provide a survey of machine learning as a tool in the econometric toolbox.

In the field of international trade, there are to the best of our knowledge only very few applications of these advanced tools. Alschner et al. (2018) introduce a database of full text corpus of 448 WTO-notified trade agreements and discuss potential uses in economics, political science, and law. Wohl and Kennedy (2018) use an NN to estimate a gravity model for international trade flows. They compare an OLS specification, a Poisson-Pseudo Maximum Likelihood specification, and an NN in terms of their accuracy for out-of-sample estimates. The NN with country fixed effects leads to the predictions with the lowest root mean squared error. Circlaeys et al. (2017) and Quimba and Barral (2018) also show the potential usefulness of machine learning algorithms for predicting international trade flows. Batarseh et al. (2019) apply various machine learning methods to investigate the most influential economic predictors of trade of specific commodities. Gopinath et al. (2020) use machine learning techniques to learn trade patterns in agricultural trade flows. They show that while supervised machine learning techniques are suitable to find out the key economic factors underlying agricultural trade flows, unsupervised approaches provide better fits over the long-term. Akman et al. (2020) study the impact of visa policies on bilateral trade for Turkey using Support Vector Machines and NNs, and find that visa restrictions have a significant impact on exports. Lahann et al. (2020) train an NN to predict the effective FTA utilization for export transfer transactions based on transaction characteristics such as import country, export country, or product type. We contribute to this literature by using machine learning algorithms to explore potential non-linearities and complex interactions in economic determinants for predicting RTA formation.

The rest of the paper is split into two parts. Sections 2 and 3 examine empirical models for RTA formation based on existing theories of RTA formation using probit models and machine learning, while section 4 discusses the evaluation of model predictions. In particular, Section 2 describes the data and construction of variables, while Section 3 provides the empirical specifications, methods and estimates of the benchmark probit model, our tree-based methods and the used NNs. Section 4 compares the performance of the probit model and the NN and gives out of sample predictions for the conclusion of RTAs of the EU and the United States with its trading partners using the probit model as well as the trained NN. The last section concludes.

2 Data

The goal in collecting the data was at the one hand to include the well established explanatory variables to explain RTA membership, and on the other hand to cover as many countries as possible. The latter is specifically important when we employ machine learning methods, which are known to work best with enough data.

For our dependent variable, regional trade agreements, we use Mario Larch's Regional Trade

Agreements Database from Egger and Larch (2008)[2]. This data covers all RTAs notified to the WTO for the last 70 years from 1950 to 2019. RTAs include CUs, FTAs, PSAs, EIAs, CU&EIAs, FTA&EIAs, and PSA&EIAs, following the classification of the WTO.

Our machine learning models use raw data for real GDP of the country-pairs (RGDP_l and RGDP_s), population (POP_l and POP_s), distance (DIST_{ls}), and whether countries belong to the same continent (DCONT_{ls}). For real GDP (RGDP) and population (POP) data we use the World Development Indicators.[3] Real GDP are in 2010 US$. We sort the GDP and population data for the respective country-pairs according to their GDP, indexing the GDP and population of the larger and smaller country by l and s, respectively. We sort country-pairs in this way in order to meaningfully interpret the results for the country-specific variables and to apply the fitted model to unseen data. The descriptive statistics for these variables are listed in Table 1. We determine geographical distances between countries, DIST_{ls}, using centroid data from the Harvard Worldmap[4], and calculate the shortest distance between the economic centers of two countries based on the longitudes and latitudes using the Haversine formula. We include in our analysis the dummy DCONT_{ls}, which takes the value one if countries l and s are on the same continent, and zero else. We distinguish between Africa, Asia, Europe, North America, Oceania, and South America. Being on the same continent is expected to increase the likelihood that two countries form an RTA.

For the probit model, we use constructed variables based on these raw data. The first constructed explanatory variable is a measure for geographical distance, NATURAL_{ls}, which is the log of the inverse of the geographical distance:

$$\text{NATURAL}_{ls} = \log\left(\frac{1}{\text{DIST}_{ls}}\right). \tag{1}$$

We expect NATURAL to have a positive influence on the probability to form an RTA.

We follow Baier and Bergstrand (2004) and Egger and Larch (2008), and use the RGDP and population data to construct the variables RGDPsum, which is the log of the sum of the RGDPs of a country pair ls for each year t, i.e.:

$$\text{RGDPsum}_{lst} = \log\left(\text{RGDP}_{lt} + \text{RGDP}_{st}\right), \tag{2}$$

and RGDPsim, which measures the GDP similarity between two trading partners, defined as:

$$\text{RGDPsim}_{lst} = \log\left(1 - \left(\frac{\text{RGDP}_{lt}}{\text{RGDP}_{lt} + \text{RGDP}_{st}}\right)^2 - \left(\frac{\text{RGDP}_{st}}{\text{RGDP}_{lt} + \text{RGDP}_{st}}\right)^2\right). \tag{3}$$

The hypothesis is that both, RGDPsum and RGDPsim, have a positive influence on the probability to conclude an RTA. With distance and country sizes, as measured by RGDPsum and RGDPsim, we therefore capture the classical explanatory variables based on the gravity framework, which is typically used to explain trade flows. But as it was shown in the seminal article by Baier and Bergstrand (2004), the same explanatory variables help to understand RTA formation.

Further, we construct a measure of capital-labor-ratio similarity between the two countries,

[2]The data are freely available at https://www.ewf.uni-bayreuth.de/de/forschung/RTA-daten/index.html.
[3]The data can be downloaded at https://databank.worldbank.org/source/world-development-indicators.
[4]The data can be downloaded at https://worldmap.harvard.edu/data/geonode:country_centroids_az8.

DKL. Due to limited capital data availability for many countries, we follow Egger and Larch (2008) and construct the measure based on the GDP per capita motivated by the high correlation between capital–labor ratios and real GDP per capita. Hence, capital-labor-ratio similarity between the two countries is defined as the absolute value of the difference between the logs of the RGDP over population, i.e.:

$$\mathrm{DKL}_{lst} = \left| \log\left(\frac{\mathrm{RGDP}_{lt}}{\mathrm{POP}_{lt}}\right) - \log\left(\frac{\mathrm{RGDP}_{st}}{\mathrm{POP}_{st}}\right) \right|. \tag{4}$$

As the capital-labor ratio was motivated to have a non-linear effect by Baier and Bergstrand (2004), we also include the square of DKL_{lst}, i.e., SQDKL_{lst}. Note that this is the only explanatory variable that also enters with its quadratic form. The expectation is that DKL_{lst} increases the probability of forming an RTA and that the effect is decreasing with larger differences, i.e., we expect the coefficient estimate of the square of DKL_{lst} to be negative.

Based on the geographical distance, we construct the variable REMOTE, which measures the remoteness of a pair of trading partners from the rest of the world. Specifically, REMOTE is constructed as follows:

$$\mathrm{REMOTE}_{ls} = 0.5 \times \left(\log\left(\frac{\sum_{k \neq l} \mathrm{DIST}_{ks}}{n_t - 1}\right) + \log\left(\frac{\sum_{k \neq s} \mathrm{DIST}_{lk}}{n_t - 1}\right) \right), \tag{5}$$

where n_t denotes the number of country-pairs in period t. If two countries are remote from other trading partners, they should have a larger incentive to integrate with each other. We therefore expect the coefficient estimate for REMOTE to be positive.

Based on the RGDP per capita, we also construct a measure for the difference in the factor endowments between the rest of the world (ROW) and a given country-pair, DROWKL, which is defined as follows:

$$\begin{aligned} \mathrm{DROWKL}_{lst} &= 0.5 \times \left| \log\left(\frac{\sum_{k \neq l} \mathrm{RGDP}_{kt}}{\sum_{k \neq l} \mathrm{POP}_{kt}}\right) - \log\left(\frac{\mathrm{RGDP}_{lt}}{\mathrm{POP}_{lt}}\right) \right| \\ &+ 0.5 \times \left| \log\left(\frac{\sum_{k \neq s} \mathrm{RGDP}_{kt}}{\sum_{k \neq s} \mathrm{POP}_{kt}}\right) - \log\left(\frac{\mathrm{RGDP}_{st}}{\mathrm{POP}_{st}}\right) \right|. \end{aligned} \tag{6}$$

As argued by Baier and Bergstrand (2004), DROWKL is expected to have a negative effect on the probability to form an RTA, as a larger (absolute) difference between the capital–labor ratios of the member countries and the ROW's capital–labor ratio is expected to increase potential trade diversion effects when forming an RTA.

For our main analysis we use a cross-section for 2013. This ensures a wide coverage of countries with data for all explanatory variables needed. As our dependent variable only contains information for a pair, we use every country-pair only once, i.e., USA-Canada and Canada-USA is added only once to the data. We end up with 19,701 unique pairs in our sample, which results from using information for 199 countries, i.e., $n_c(n_c - 1)/2$, with n_c denoting the number of countries. This coverage is substantially larger than in previous studies. For example, Baier and Bergstrand (2004) only used information for 54 countries leading to 1,431 unique country-pairs and Egger and Larch (2008) for 146 countries, leading to 10,585 unique country-pairs. The larger coverage of countries leads to the inclusion of smaller and more remote countries

that may react differently to the explanatory factors. Besides needing a larger dataset to learn non-linearities and complex interactions between explanatory variables with machine learning algorithms, the extension of the dataset may lead to a larger heterogeneity which is potentially less of a problem in a smaller, more coherent dataset. As RTAs are concluded among most of the countries in the world (in our dataset, only 34 countries did not have any RTA with another country in the year 2013), we view the inclusion of many countries as substantial for the understanding of RTA formation.

Table 1 provides the descriptive statistics for the dependent variable and all our explanatory variables. We see that in our dataset we only have 19 % of observations with an RTA in place. Hence, overall and despite the huge surge over the last decades, having an RTA with a trading partner is still a relatively rare event. This will be important for the prediction performance, both of our probit model, as well as for the machine learning algorithms. The other explanatory variables show substantial variation. Most notably, even though our dataset is considerably larger, the descriptive statistics are qualitatively identical and quantitatively very similar to the values reported in Egger and Larch (2008). Hence, our results from the probit specifications should be comparable to previous findings.

Table 1: Descriptive statistics of cross-section data for the year 2013.

Statistic	N	Mean	Std. Dev.	Min	Pctl(25)	Pctl(75)	Max
RTA_{ls}	19,701	0.191	0.393	0	0	0	1
$RGDP_{l}$	19,701	$6.688E^{11}$	$1.898E^{12}$	$1.630E^{8}$	$2.678E^{10}$	$4.385E^{11}$	$1.585E^{13}$
$RGDP_{s}$	19,701	$4.539E^{10}$	$1.735E^{11}$	$8.125E^{7}$	$1.763E^{9}$	$2.678E^{10}$	$7.751E^{12}$
POP_{l}	19,701	$6.073E^{7}$	$1.862E^{8}$	$1.761E^{4}$	$5.080E^{6}$	$4.552E^{7}$	$1.357E^{9}$
POP_{s}	19,701	$1.045E^{7}$	$3.494E^{7}$	$1.082E^{4}$	$2.566E^{5}$	$1.005E^{7}$	$1.357E^{9}$
$DIST_{ls}$	19,701	8075.80	$4.548E^{3}$	53.76	4450.63	11342.09	19795.20
$DCONT_{ls}$	19,701	0.130	0.336	0	0	0	1
$NATURAL_{ls}$	19,701	−8.774	0.766	−9.893	−9.336	−8.401	−3.984
$RGDPsum_{ls}$	19,701	25.617	1.904	19.314	24.216	26.881	30.792
$RGDPsim_{ls}$	19,701	−2.367	1.708	−11.488	−3.262	−0.984	−0.693
DKL_{ls}	19,701	1.713	1.217	0.00005	0.713	2.504	6.564
$SQDKL_{ls}$	19,701	4.413	5.460	0.000	0.508	6.269	43.081
$REMOTE_{ls}$	19,701	8.971	0.151	8.720	8.846	9.073	9.518
$DROWKL_{ls}$	19,701	1.299	0.616	0.009	0.833	1.712	3.552

Notes: This table reports for each variable the number of observations, the mean, the standard deviation (Std. Dev.), the minimum (Min), the 25th and 75th percentiles (Pctl(25) and Pctl(75)), and the maximum (Max). E referrers to "10 to the power of". We measure GDP in 2010 US Dollar and $DIST_{ls}$ in kilometers.

3 Empirical Specifications and Methods

We start our empirical analysis of RTA membership by specifying a standard probit model. This allows us to compare results from previous findings with our data. Afterwards, we will use different tree-based methods as well as an NN to predict RTA membership, which allow for non-linearities and complex interactions between explanatory variables and do not require to

specify concrete functional forms as the relationships are learned form the data. NNs are one of the most commonly employed and flexible machine learning tools.

3.1 Probit

The specification of our probit model follows the standard in the literature (see Baier and Bergstrand, 2004; Egger and Larch, 2008, for examples):

$$\mathbf{RTA}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{7}$$

where $\mathbf{RTA}^*$ is the vector of size $N \times 1$ (vectors and matrices are in bold) of the unobservable utility differential from membership relative to non-membership in an RTA for all country-pairs, $\mathbf{X}$ collects all K explanatory variables and is of size $N \times K$, $\boldsymbol{\beta}$ is a unknown parameter vector of size $K \times 1$, and $\boldsymbol{\epsilon}$ is an error term assumed to be independent of $\mathbf{X}$ and standard normally distributed.

The unobservable $\mathbf{RTA}^*$ is translated to the vector of observable RTA membership $\mathbf{RTA}$ by:

$$\mathbf{RTA} = \mathbf{1}[\mathbf{RTA}^* > 0], \tag{8}$$

where $\mathbf{1}[\cdot]$ denotes an indicator function taking value 1 if two countries have an RTA (when $\mathbf{RTA}^* > 0$), and 0 otherwise (in which case $\mathbf{RTA}^* \leq 0$). The conditional probabilities, $Pr[\cdot|\mathbf{X}]$, in which we are interested, are given by:

$$Pr(\mathbf{RTA} = 1|\mathbf{X}) = Pr(\mathbf{RTA}^* > 0|\mathbf{X}) = \Phi\left(\mathbf{X}\boldsymbol{\beta}\right), \tag{9}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, which ensures that $Pr(\mathbf{RTA} = 1|\mathbf{X})$ is between 0 and 1.

Note that while the probit model is a non-linear model, the data and parameters enter the non-linear mean function $\Phi(\cdot)$ by the single index $\mathbf{X}\boldsymbol{\beta}$, i.e., it is a single-index model. This is what Cameron and Trivedi (2005) call non-linearity of the mild form. It has the advantages that the point estimates can be interpreted in terms of the relative effects of changes in regressors, and, because $\Phi(\cdot)$ is monotonic, the signs of the estimated coefficients are identical to the signs of the marginal effects. The disadvantage is that while the mean is a non-linear function, the regressors and parameters are linear combinations, which severely limits the type of non-linearity the model can capture.

Table 2 presents our benchmark estimation results of the probit model. Column (1) uses as explanatory variables distance and the size terms. As can be seen, all of them are highly significant and have the expected signs. Column (2) adds a measure of capital-labor-ratio similarity DKL. Following Baier and Bergstrand (2004) and Egger and Larch (2008), we also add the square term SQDKL. This is specifically interesting for us, as it is the only direct non-linear explanatory variable in the single-index model. As expected from theory, DKL is increasing the probability to sign an RTA, while the effect is decreasing with increasing dissimilarity. Column (3) adds several additional controls, such as DCONT, REMOTE, and DROWKL. All of them are highly significant and have the expected signs.[5] The explanatory power of the regressions

[5]Note that DROWKL is significant but positive for all RTAs in the cross-section in Egger and Larch (2008).

Table 2: Probit results for the probability of forming an RTA based on the year 2013.

		Dependent variable:			
	Theory	RTA			
		(1)	(2)	(3)	(4)
NATURAL	(+)	0.541^{***}	0.529^{***}	0.474^{***}	0.718^{***}
		(0.014)	(0.014)	(0.018)	(0.024)
RGDPsum	(+)	0.231^{***}	0.234^{***}	0.216^{***}	
		(0.007)	(0.007)	(0.007)	
RGDPsim	(+)	0.164^{***}	0.145^{***}	0.141^{***}	0.027^{*}
		(0.008)	(0.008)	(0.008)	(0.012)
DKL	(+)		0.090^{***}	0.068^{*}	0.098^{*}
			(0.035)	(0.035)	(0.044)
SQDKL	(-)		-0.084^{***}	-0.064^{***}	-0.106^{***}
			(0.009)	(0.009)	(0.012)
DCONT	(+)			0.368^{***}	0.469^{***}
				(0.034)	(0.050)
REMOTE	(+)			0.183^{**}	
				(0.087)	
DROWKL	(-)			-0.201^{***}	
				(0.022)	
Constant		-1.792^{***}	-1.839^{***}	-3.368^{***}	
		(0.206)	(0.208)	(0.784)	
Observations		19,701	19,701	19,701	16,727
Pseudo R^2		0.152	0.183	0.192	0.367
Log Likelihood		$-8{,}152.293$	$-7{,}852.044$	$-7{,}764.139$	$-5{,}295.22$
Akaike Inf. Crit.		16,312.590	15,716.090	15,546.280	-

Notes: This table reports probit estimates using data for a cross-section of 2013. The first column includes gravity controls, the second column adds our measure for capital-labor-similarity, the third column adds further controls, and the last column includes exporter and importer fixed effects. *p<0.1; **p<0.05; ***p<0.01. See text for further details.

Table 3: Prediction performance using a decision tree.

	Actual 0	Actual 1
Predicted 0	0.629	0.237
Predicted 1	0.371	0.763

is a bit below 20% as measured by McFadden's R^2.

In column (4) of Table 2 we allow for exporter and importer fixed effects. Only recent developments allow to estimate generalized linear models with multiple fixed effects in comparably large datasets, see Stammann (2017) and Bergé (2018). We use the `fixest`-package[6] in R to estimate our probit model with importer and exporter fixed effects. When using exporter and importer fixed effects, RGDPsum and RGDPsim are perfectly collinear, hence, we only include RGDPsim capturing more meaningful non-linearity. Similarly, DROWKL is perfectly multicollinear with the exporter and importer fixed effects.[7] REMOTE is very highly correlated with NATURAL when controlling for exporter and importer fixed effects.[8] Hence, we also exclude it from our regression. The results from column (4) show that all the coefficients from the remaining variables have the expected signs, and NATURAL, SQDKL and DCONT are also highly significant. The R^2 increases to about 37 %. Given the substantial increase in the explanatory power, our main and preferred specification is the one in column (4). Furthermore, table 3 provides an overview of the predictive performance in terms of false positives and false negatives, also referred to as *insufficient bilateralism* and *excessive bilateralism* (see Baier and Bergstrand, 2004). This can be interpreted as the pairs for which the model predicted an RTA despite there not being one or analogously pairs which should not have an RTA according to the model but which do however have an agreement. In a sense, the objective is to minimize the number of cases of miss-classification or excessive and insufficient bilateralism in order to find a model representation that best fits the observed data.

Overall, we view our benchmark results for the probit model encouraging as they support the predictions from the theory, show explanatory power and are in line with previous findings (even though with substantial wider coverage of countries, adding considerable heterogeneity). However, while we see that the explanatory variables are highly significant and have explanatory power, we still see substantial variation of RTA membership unexplained. While the probit model seems to capture the general trend of the considered variables, it is not able to capture more intricate relationships. Furthermore, it is vulnerable to miss-specification. Due to this we continue on to tree-based methods and NNs.

3.2 Tree-based Methods

We now turn to a decision tree to learn about the important predictors (i.e., explanatory variables) and their non-linear impact on predicting RTA membership. Decision trees can be used for regressions, i.e., when the dependent variable is a quantitative variable, such as trade flows, and classification problems, i.e., when the dependent variable is a qualitative response, such as in our case RTA membership (which is either yes (=1) or no (=0)). Decision trees do not make any functional form assumptions; they are completely non-parametric. Hence, they do not make assumptions on the data generating process besides independence between observations (see Taddy, 2019). Rather, they split (stratify, segment) the predictor space (i.e., the space spanned by the explanatory variables) into a number of regions. All observations that are in

Hence, using more recent data and data for more countries seems to bring the probit estimates closer to the theory.

[6] Available for download at https://CRAN.R-project.org/package=fixest.

[7] We provide formal demonstrations of these multi-collinearities in the Online Appendix.

[8] A regression of REMOTE on NATURAL and importer and exporter fixed effects delivers and R^2 from basically 1 and a residual standard error of 0.0008558.

the same region are then predicted to have the same response. For regression trees, the mean or mode of the observations falling in the same region is used as predicted response, and for classification trees the most commonly occurring class of the observations falling in the same region is used as predicted response.

Tree-based methods are convenient in that they are very intuitive and that they can be nicely displayed graphically. This is a big advantage, as many machine learning algorithms occur as black boxes as the actual fitted model is hard to grasp, which could be one reason why they are not (yet) frequently used in some fields, such as international trade. Using this advantage of decision tress, we plot a decision tree for our RTA formation in order to highlight the underlying non-linearity the decision tree unveils. However, one has to keep in mind that (simple) decision trees are typically outperformed by other machine learning methods in terms of predictive power (see James et al., 2013). The reason is that non-parametric methods, such as decision trees, easily overfit due to their flexibility. Additionally, classical regularization techniques, such as penalized-deviance or cross validation selection, do not work for non-parametric methods as they are sensible to small changes in the data. One can use bootstrap aggregating ("*bagging*"), which is a special case of random forests. With bagging one runs multiple with-replacement samples and then uses the mean fit across bootstraps as an estimate for the average model fit. This will help if the number of explanatory variables relative to the number of observations is comparably low. However, going from the tree to the forest leads to a loss of the possibility to graphically illustrate the results. As we also employ fixed effects, and to balance stability and dimension reduction, we therefore will also employ NNs as a semi-parametric method and compare them in terms of their predictive power with the probit model and decision trees.

We implemented the decision tree algorithm using the `rpart`-package[9] in R. In particular, we use a greedy splitting rule, meaning that the predictor space is partitioned according to the predictor which yields the greatest improvement in accuracy for the prediction. The increase in accuracy is determined by the Gini index, which is defined by:

$$G = \sum_{m=1}^{M} \hat{p}_{km} \left(1 - \hat{p}_{km}\right), \tag{10}$$

where $\hat{p}_{km}$ is the proportion of training observations in the kth region that are of class m (James et al., 2013). When setting the minimum number of observations in a node for a split to be attempted to five, the smallest size of a leaf to one and the complexity parameter which determines the required increase in the fit for a split to be made at 0.0004, we obtain a prediction performance on the test data as seen in Table 4. As can be seen, 83 % of country-pairs with no-RTA are correctly predicted, and 85 % of country-pairs with RTA are correctly predicted.

Table 4: Prediction performance using a decision tree.

	Actual 0	Actual 1
Predicted 0	0.829	0.147
Predicted 1	0.171	0.853

We did not employ pruning techniques on the tree, which would reduce the number of nodes

[9] Available for download at https://cran.r-project.org/package=rpart.

and leaves to reduce over-fitting, yet setting the complexity parameter already indicates which splits would be omitted in the pruning process.

Figure 5 plots the outcome of a simpler decision tree for which the complexity parameter is set to 0.003. Note that due to the aforementioned greedy algorithm, these splits also occur in a more complex version of the tree. Although it is not possible to obtain point estimates for tree-based methods, as usually the case in econometric inquiry, we are able to interpret the resulting tree, which given its superior performance when comparing tables 3 and 4, seems to be a better representational model of the data generating process than the probit. The first node is the starting point which contains all observations (explaining the 100 %), showing that RTA is equal to one for 18 % of the observations given our training data. In line with the outcomes of the probit model, distance is an important predictor. Specifically, the first split is based on a distance above or below 3,096 kilometers. Any pair of countries below this distance is already reaching its terminal node. 15 % of the observations fall in this category. The prediction for these pairs is that there is an RTA in place, which is actually the case for 47 % of the observations in this terminal node. If the distance is larger than 3,096 kilometers, the next node splits the remaining 85 % observations according to the GDP of the poorer country of the pair. If the GDP is lower than 42bn $, the next internal node splits the observations according to population of the poorer country. Then again, distance appears, and so on. Until we end up at a terminal node. There are 25 terminal nodes, ten of which predict an RTA in place, while the rest predict that there is no RTA in place. Most importantly, distance, population and GDP appear in different internal nodes with different values where they are split. For example, while the first split is at a distance of 3,096 kilometers, we see other distance splits at 5,748, 7,020, 7,896, 9,186, 10,000 and 15,000 kilometers. This shows that the decision tree uses distance very non-linearly when predicting RTAs. The same is true for population and GDP. Overall, it can be seen that conventional determinants of regional trade do play a major role in the formation of RTAs, which is in line with economic theory. However, it is also apparent that these relationships are not as simple and linear as they have been empirically specified so far.

As shown in Table 4, a regular decision tree is already quite capable in navigating complex environments. Yet, due to the greedy nature of the algorithm it is possible to perform splits that might yield the highest gain in accuracy at that point during the training but are not advantageous when considering later splits, which can result in over-fitting on the training data. This is especially pronounced when considering high dimensional predictor spaces. Due do this we introduce a random forest, using the `randomForest`-package[10] in R, as a more advanced tree-based method. In particular, a random forest works by fitting many different decision trees each using a different, random subset of the original set of the predictor space at each split. When predicting the implied probability of a new observation the random forest averages over all of the individual trees that were grown during the training process (James et al., 2013). This, however, comes at the cost of interpretability as the resulting decision of a fitted model can no longer be visualized as shown for the original tree used for Figure 5. Given our data we fitted 100 different trees, where each tree had $\lceil \sqrt{K} \rceil = \lceil \sqrt{6} \rceil = 3$ predictors randomly chosen for each split (see James et al., 2013).[11] In order to compare the performance, we use the ROC (=Receiver Operating Characteristic) curve. The ROC curve plots the rate of false positives

[10] Available for download at https://cran.r-project.org/web/packages/randomForest/.

[11] $\lceil \cdot \rceil$ denotes the ceiling function, which always rounds to the next largest integer.

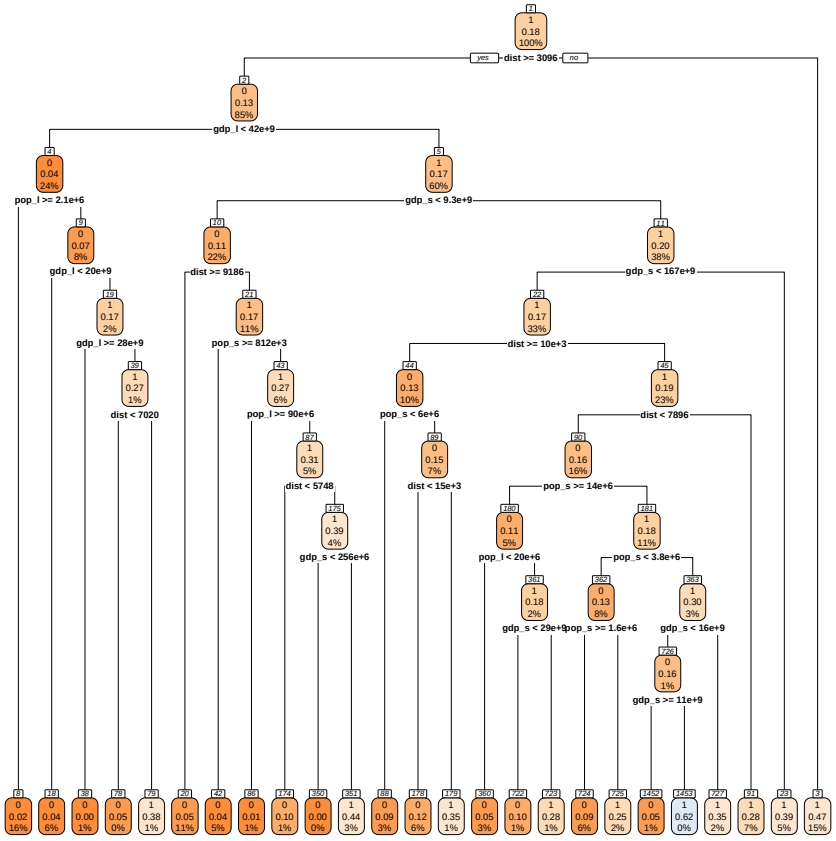

Figure 5: Simplified decision tree.

(=the model predicts an RTA where there is none, insufficient bilateralism) against the rate of true positives (=the model predicts an actual RTA) for all possible thresholds (see Hastie et al., 2009). The threshold is the cut-off probability that has to be exceeded in order to classify the prediction as being equal to one. The best outcome a model can achieve will lead to a ROC curve that hugs the top left corner, implying that we have a high true positive rate and a low false positive rate. In the case that our exogenous variables do not have explanatory power, the ROC curve will be a 45 degree line. As can be seen from Figures 6 and 7 both methods perform rather well in our base case. Nonetheless it can be seen in the former figure that the random forest seems to outperform the decision tree on unseen data, which is in line with the already outlined tendency of the regular tree to overfit.

Another approach to improve the prediction performance of decision trees is boosting. In contrast to random forests that use random subsets to fit different decision trees, gradient boosting machines grow trees sequentially. Hence, each tree that is grown uses information from previously grown trees (James et al., 2013). This improves the performance, as each newly trained model specializes in addressing the weak points of the previous models. Gradient boosting machines are well suited for problems where structured data, as in our case of RTA predictions, is available (see Chollet and Allaire, 2018, for example). We implement the gradient boosting machine using the `xgboost`-package in R.[12] In order to hyperparameterize the gradient boosting machine, we use a genetic algorithm from the `GA`-package (see Scrucca, 2017) which acts as an efficient search algorithm for finding optimal combinations of parameters for the task at hand in a large search space and thus reduces the ambiguity in the hyperparamter tuning process.[13] Specifically, we span a high dimensional fitness function whose objective is the Area Under Curve (AUC), i.e., the area between the ROC curve and the 45 degree line from the origin ranging between 0 and 1. The AUC is based on the validation data over which we optimize for the maximum number of gradient descent iterations (allowing values between 10 and 1000), the minimum loss reduction required to make a further partition on a leaf node of the tree (γ, allowing values between 5 and 100), control the balance of positive and negative weights (allowing values between 5 and 100), maximum depth of a tree (allowing values between 4 and 8), minimum number of instances needed after which no further split is attempted and the node thus becomes a leaf (allowing values between 2 and 5), and a parameter controlling the learning rate (η, allowing values between 0.1 and 0.5) which determines the magnitude of the gradient update in each iteration. We train the model on a training dataset (which is a random draw of 70 % of the observations) from which we hold out 1000 observations as validation dataset on which we validate the hyperparametrization, before we test the optimized gradient boosting machine on the remaining 30 % of the observations used as test dataset. We report in Table 5 the set of optimal hyperparameters resulting from the GA-algorithm. The evolution of the AUC over the generations of optimization for the gradient boosting machine without and with fixed effects are shown in Figures 24 and 25 in the Appendix. Both plots suggest that the best model was found in the given number of generations, which can be inferred form the fact that the best performing genotype incurs no changes for a long time. Comparing the results from

[12] Available for download at https://cran.r-project.org/web/packages/xgboost/.

[13] Available for download at https://cran.r-project.org/web/packages/GA/index.html. Alternative ways to do automatic hyperparametrization optimization are Bayesian optimization or simple random search (see Chollet and Allaire, 2018, for exampe).

the optimized gradient boosting machine with the results form the decision tree and the random forest by looking at Figures 6 and 7, we see that the gradient boosting machine outperforms both. The ROC curve is even further to the top-left and Figure 7 shows that the gradient boosting machine manages to obtain over 87.5 % true positives and true negatives at the same time. It seems to balance flexibility to fit the data and the risk of overfitting the training data best.

Table 5: Hyperparametrization XGBoost.

Hyperparameter	without FE	with FE
Iterations	780	847
γ	17.8	19.8
Scale of positive weight	53.4	54.5
Depth	6	5
Minimum child weight	4.2	3.5
η	0.25	0.24

Notes: "FE" is the abbreviation for Fixed Effects.

A nice feature of gradient boosting machine is that they allow to investigate the relative relevance of the different explanatory variables. Figure 8 fits a gradient boosting machine for each year from 1960 to 2018. There are a number of interesting findings: i) Distance is the most relevant factor, specifically in the early years. ii) In the mid-1960s we see a fall in the relevance of distance. At the same time the influence of two countries belonging to the same continent increases. This maybe explained by the Arab Common Market entering into force in 1965 and the Australia and New Zealand FTA getting effective in 1965. iii) The decreasing importance of distance in 1973 can be understood by EU enlargements and various agreements that the EU concluded (Switzerland, Liechtenstein, Iceland, Norway) at that time as well as the Protocol on Trade Negotiations that covered countries that are far apart. This heterogeneity diminished the explanatory power of distance. At the same time, large and small countries thereby formed an agreement, leading to a lager explanatory power of GDPs, specifically of the smaller country. iv) Another large drop of the influence of distance can be seen in 1989/1990. In 1989 the Global System of Trade Preferences among Developing Countries (GSTP) came into force, which covered many countries with large distances, as for example Chile, Egypt, India, and Iraq. While this reduces the predictive power of distance, the fact that the GSTP covered developed countries increases the importance of GDP as explanatory factor. v) From 1990 onward the importance of distance increases again, mostly at the cost of being on the same continent. RTAs are no longer bound to continents. Overall, the importance of GDP and population also stay high or increase, which also reflects that in later periods more RTAs are in place.[14]

Overall, we see that tree-based methods do a good job in predicting RTA formation, which is a comparably structured problem. Given the comparison of the performance of the different tree-based methods and the fact that we will increase the predictor space in further analysis, we proceed further analysis using the gradient boosting machine as our tree-based method of choice.

[14]Note that these are only possible explanations. It would be interesting in future research to dig deeper into the specific events and mechanisms.

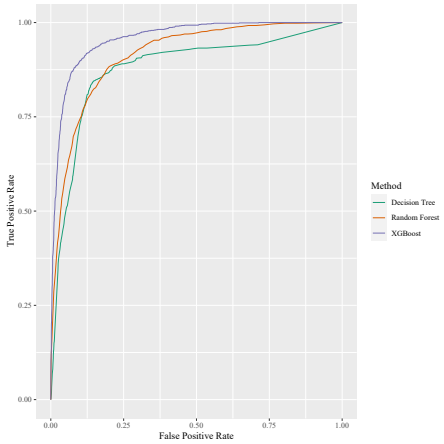

Figure 6: Comparing all tree based methods using a ROC curve.

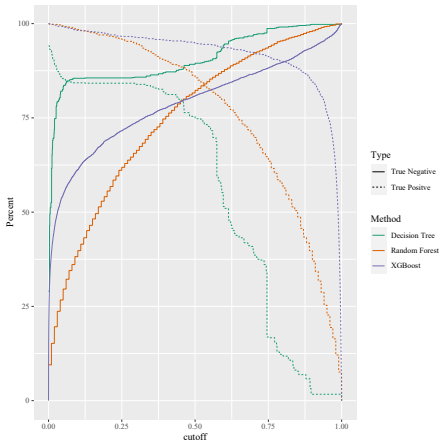

Figure 7: Comparing all tree-based methods in their sensitivity to different cut-off values.

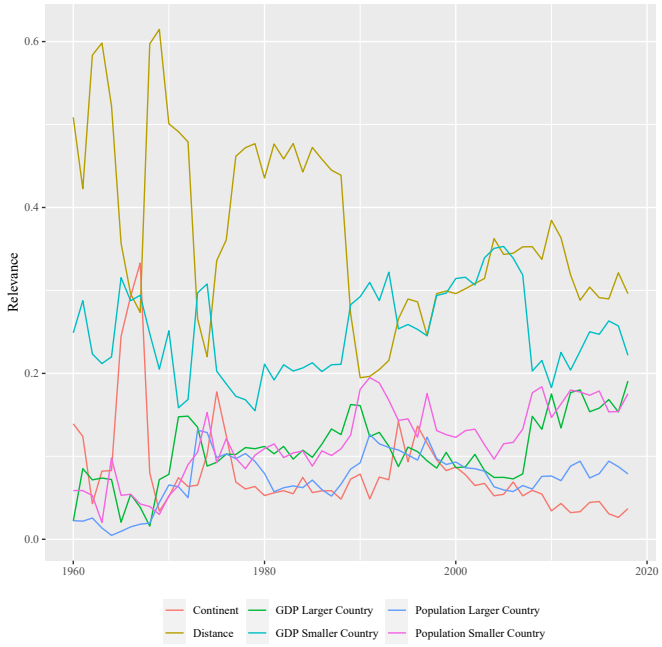

Figure 8: Relative relevance in predicting an RTA from 1960 to 2018

3.3 Neural Network

For our probit specification, we had to make strong functional form assumptions. Most importantly, we assumed that the data and parameters enter the non-linear mean function by the single index $\mathbf{X}\boldsymbol{\beta}$. This imposed strong functional form restrictions are more or less arbitrary and may be very far from the underlying data generating process. Very likely, as motivated by our Figures 3 and 4, they are too simplistic and are not able to capture the fundamental non-linearity. We therefore also introduced tree-based methods, which are fully non-parametric. This flexibility comes at the risk of overfitting the training data, specifically with many predictors. In order to deal with both of these shortcomings we introduce NNs which turn out to be rather adept at mastering the trade-off between over- and under-fitting the data.

In the mid 1980s NNs were introduced. NNs are highly non-linear, semi-parametric models that offer a wide array of functional forms that do not have to be specified in advance, thus offering much more flexibility than a probit specification in fitting any smooth relationship, yet putting some parametric structure. This flexibility comes at the cost of transparency in the interpretation of the model results as there is no easy way, similar to that of for example linear and single-index models, to interpret the weighting matrix resulting from the training process (see Jahn, 2018). Nonetheless, they are particularly adept at modeling the highly non-linear processes that underlie our world making them superior to linear and single-index models with regards to the representation of data generating systems (see Herbrich et al., 2001; Storm et al., 2019). Consequently, this removes some of the vagueness that surrounds the model building process in classical methods. Another advantage of NNs over linear and single-index models as well as tree-based models is their ability to cope with collinearity in the predictor space, a problem which we encountered in our probit specification. This is due to the overparametrization and the corresponding redundancy of single inputs in the network that renders individual weights insignificant (see De Veaux and Ungar, 1994). Given that multi-collinearity is a frequently encountered problem in econometric research, NNs are primed to be adapted in empirical modeling.

An NN consists of several layers, starting with an input layer that contains the explanatory variables, each in one so called neuron or node. These are followed by one or several hidden layers, that non-linearly transform linear combinations of the inputs using so called activation functions from the neurons of the previous layer (each node of a layer using different weights to sum the inputs from the previous layer). Finally an output layer, which in our case contains the predictions. If there is no non-linearity in the hidden layers, the NN reduces to a generalized linear model. But with non-linear transformations in the hidden layers, NNs can fit highly non-linear smooth relationships. In the 1990s complexity was added to NNs by allowing for more and more nodes/neurons within layers (i.e., adding width). While this allowed to learn complex relationships with enough data, it turned out to be inefficient. Recognizing this, NNs went a bit out of fashion, before they gained momentum again after 2010 under the term "deep learning". The main innovation of deep NNs is that they contain more than one, sometimes many, hidden layers. This increased the efficiency of NNs and makes them to one of the most heavily used machine-learning method in many situations nowadays. NNs are typically estimated via Maximum Likelihood, using a variety of regularizations.

We constructed the NNs using the `Tensorflow`[15] and `Keras`-packages[16] in R. In a similar fashion to the gradient boosting machine, we use a genetic algorithm for the hyperparameter tuning of the NN. Specifically, we optimize the hyperparameters for the drop-out rate in each layer (allowing values between 0 and 1), a parameter controlling the learning rate (allowing values between 0.001 and 0.01), the number of epochs to train the model (allowing values between 100 and 800), the batch size (allowing values of 32, 64, 128, 256, 512, and 1024), the number of hidden layers (allowing values 1, 2, 3, 4, and 5), control the balance of positive and negative realizations of our dependent variable (allowing values between 1 and 10), and a parameter for each layer that determines the number of neurons/units (allowing values 8, 16, 32, and 64). The tuning process follows the exact same rules as already outlined for the gradient boosting machine. Figures 26 and 27 in the Appendix show the evolution of the AUC without and with fixed effects, respectively. Similar to the gradient boosting machine, in both plots the best genotype ceases to improve at the end of the generations shown, which suggest that the best model was found. Our optimized NN consists of one input layer taking in RGDP_l, RGDP_s, POP_l, POP_s, DIST_{ls} and DCONT_{ls}, three hidden layers $\{h_1, h_2, h_3\}$ and one output layer, which returns $Pr(\text{RTA} = 1)$. The results of this process can be found in Table 6, where the learning rate refers again to the velocity of gradient descent and weight to the cost for miss-classifying a 1 as a 0 in order to deal with the imbalance in the number of RTAs in the data as reported in Table 1.

Table 6: Hyperparametrization NN.

Hyperparameter	without FE	with FE
Drop out	0.212	0.219
Learning rate	0.0045	0.0039
Epochs	444	20
Batchsize	256	256
Number of hidden layers	3	3
Weight	5	6.6
Neurons per hidden layer	32, 32, 16	64, 32, 16

Notes: "FE" is the abbreviation for Fixed Effects.

All neurons are structured in a dense feed forward setup, meaning that each neuron in a layer has an input from each neuron in the preceding layer. Each hidden layer is followed by a dropout layer, which randomly sets a given share of outputs from the previous layer to zero (21.2 % in our optimized NN) in order to avoid over-fitting on the training sample (see Srivastava et al., 2014). The activation function—the non-linear transformation in each neuron—is the *Rectifier* with the functional form $f(x) = max\{x, 0\}$.[17] The output layer uses a *Softmax* function taking the form $\sigma_i(z) = e^{z_i}/(\sum_{j=1}^{K} e^{z_j})$ with $i = 1, .., K$ being the inputs from the preceding layer. This last transformation is needed to ensure that the outputs represent probabilities, similar to the cumulative distribution function used in the standard probit specification.

As in the case of tree-based methods, NNs are also not straightforwardly suited for interpre-

[15] Available for download at https://cran.r-project.org/package=tensorflow.

[16] Available for download at https://cran.r-project.org/package=keras.

[17] While in the 1990s a big focus was on investigating the effect of different activation functions, the consensus nowadays is that a simple, computationally efficient non-linear transformation is sufficient if one uses enough nodes and layers (see Taddy, 2019).

tation of coefficients. However, the patterns that are being detected by the NN can be studied. Admittedly, this is not as precise as the numerical values of a point estimates, yet given the predictive superiority, this loss of precision seems warranted. The information compression in readily interpretable coefficients comes at the cost of being worse in terms of description of complex data generating processes. Hence, a less explicit model may provides a better balance between precision and flexibility. A way of providing a more transparent and tangible decision process, as well as probing conventional assumptions regarding economic theory, is to apply the trained network to simulated data in order to investigate how small changes in inputs will influence the predicted outcome. Figures 9 and 10 illustrate this by holding POP_l, POP_s and DIST_{ls} constant at the mean and varying only the respective real GDP's. Note that the axes are in terms of normalized RGDP. This is important when interpreting the figures as one is looking at relative sizes, meaning that direct comparisons in real GDP between countries on the x and y axis are not feasible. The figures emphasize the non-linear nature of the decision process as the depicted shapes clearly follow some complex rules practically impossible to map in classical models. Figures 9 and 10 distinguish between countries on the same and different continents, respectively, which is one important proxy used by Baier and Bergstrand (2004) to distinguish between "natural" and "unnatural" trading partners. Besides the high non-linearity of real GDP interactions between countries, the figures also show substantial differences between "natural" and "unnatural" trading partners. Specifically, for "unnatural" trading partners RTAs are mainly predicted if both countries are comparably large.

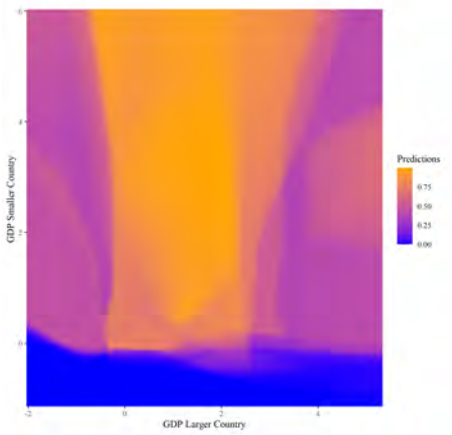

Figure 9: Decision pattern of the trained NN given that the two trading partners are on the same continent.

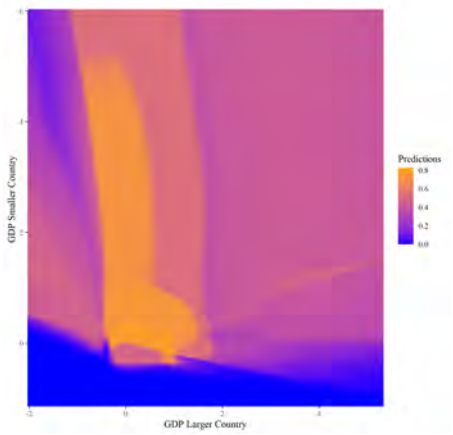

Figure 10: Decision pattern of the trained NN given that the two trading partners are on different continents.

Similarly, Figures 11 and 12 hold constant RGDP_l, RGDP_s, POP_l, and POP_s and show how DIST_{ls} influences the predicted outcome. Again, we distinguish between "natural" and "unnatural" trading partners. For countries on the same continent (Figure 11) we see a clear drop of the predicted probability to form an RTA followed by a slight increase before a steep drop off, the relationship between countries on different continents is highly non-linear in distance. This could be due to fixed costs of transportation that may also result from mode-switching (see Hummels and Schaur, 2013; Bernhofen et al., 2016), or different transport infrastructure and economic centers, which a flexible functional form for trade costs can capture.

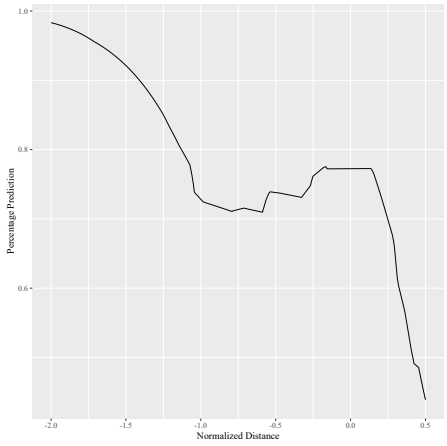

Figure 11: Implied probability of RTA formation as a function of distance for countries on the same continent.

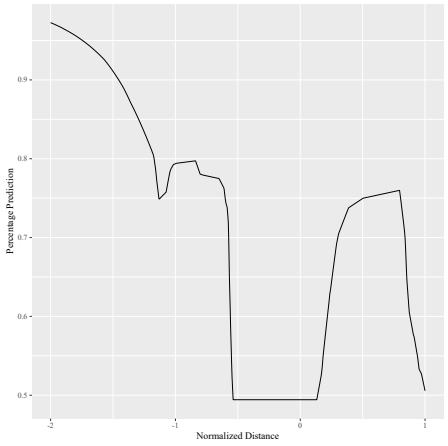

Figure 12: Implied probability of RTA formation as a function of distance for countries on different continents.

While we see from these results that both, the tree-based methods and the NN, capture substantial non-linearities, we will next investigate whether this helps to improve the predictions of RTA formation.

When it comes to theory building in supervised machine learning, the process is decidedly different from traditional econometric modeling aside from the first step, namely the selection of predictors to be employed. Afterwards machine learning methods are free to build a model representation given the data that provides the best fit. It is hence much less driven by the preferences and capabilities of the scholar but more by ideals of performance and accurate description of processes. It thus also is much less prone to miss-specification and able to search a much wider space of possible hypotheses than conventional methods. Given the comparison of predictive performance of the two approaches we find strong evidence in favor of more flexible methods when trying to find an accurate description of the data. We do however also recognize the drawbacks of said methods.

4 Comparison of Prediction Performances

After training our three different representations on the data, the probit model, the gradient boosting machine, and the NN on 70 % of the same data, we then applied all three models to the hold-out 30 % test data in order to compare the prediction performance of the three models. We will first compare the general prediction performance of the models, and then compare the prediction performances for specific events. This exercise is complementary to theory testing in the previous section, although still related to it, as inference and prediction are two different objectives of empirical analysis.

4.1 General Prediction Performance

In order to compare the general prediction performance, we use several measures. First, we employ the ROC curve that plots the false positive rate (=the model predicts an RTA where

there is none, insufficient bilateralism) against the true positive rate (=the model predicts an RTA that actually exists) for all possible cut-off levels. Figure 13 shows the outcomes of the three types of models. For each model we show the outcome for the case with and without exporter and importer fixed effects. There are three noteworthy findings: i) The ROC curves with fixed effects are above the one without for the probit and the NN. The ROC curve of the gradient boosting machine with fixed effects is very similar to the one without fixed effects. This could be due to the fact that tree-based methods are a non-parametric which may overtrain when faced with a lot of parameters. Due to the outperformance of the fixed effects model, we will mainly rely on our fixed effects models in the subsequent whenever possible. ii) In terms of the relative comparison of the three models, we find that the NN and the gradient boosting machine vastly outperform the probit model in predictive classification capability. iii) The comparison of the NN with the gradient boosting machine shows that the gradient boosting machines outperforms the NN without fixed effects, but the NN outperforms the gradient boosting machine with fixed effects. Overall, we can conclude from Figure 13 that for any cut-off level the NN with fixed effects beats the probit model and the gradient boosting machine.

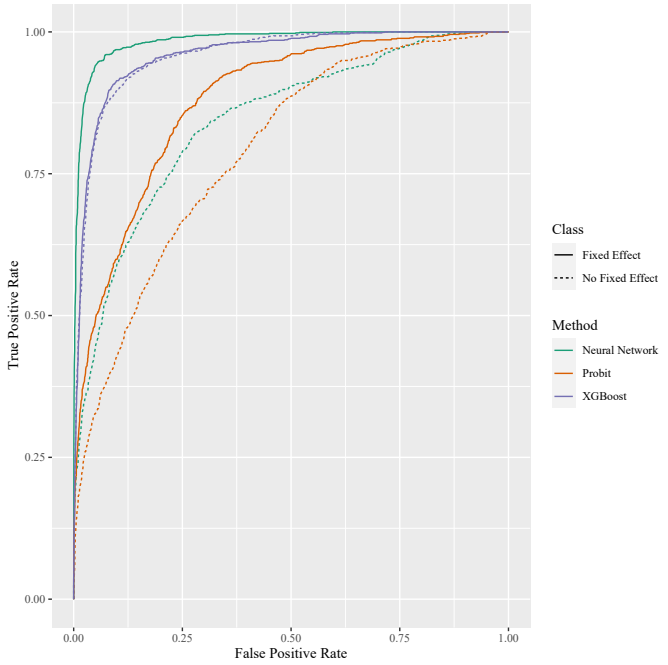

Figure 13: Comparison of the three methods using ROC curves in models with and without exporter and importer fixed effects.

Figure 14 shows the trade-off between the true positives and true negatives for all three models with fixed effects as a function of the cut-off, determining when the implied probability predicts an RTA in place.[18] For a cut-off of zero, all models predict an RTA in place for instances, implying that the true positives are predicted with a probability of one and the true negatives with a probability of zero. With increasing cut-off values, the number of correctly predicted RTAs (=true positives) decreases, while the number of correctly predicted cases of no-RTA (=true negatives) increases. Figure 14 shows that the curves for the true negatives is concave

[18]The corresponding figure without fixed effects is in the Appenidx as Figure 23.

for all three models, while the curves for the true positives is only concave for the NN and the gradient boosting machine, but not for the probit model. This implies that the share of true positives falls comparably quickly with the cut-off for the probit model, leading to a lower value of "optimal" cut-off to balance the true positives and true negatives. Furthermore, this means that the probit strongly reacts to changes in the cut-off level which follows from its tendency to produce probabilities around around the middle of the distribution, unlike the other two models that tend to the extremes (closer to 0 or 1), hence being more "certain" of their forecast. It can also clearly be seen that the probit is outperformed by both, the gradient boosting machine and the NN, and that the NN is able to obtain a true positive and true negative rate of over 90 % at the same time.

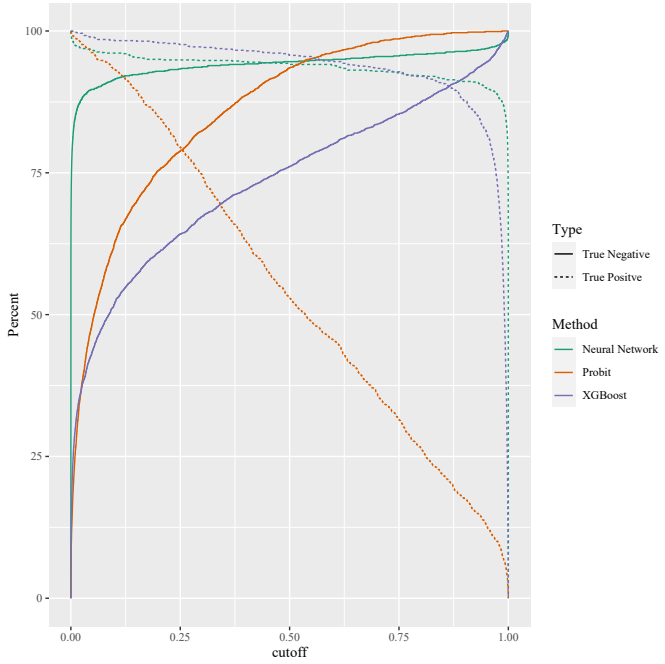

Figure 14: Trade off of true positives and negatives depending on the cut-off value at which the implied probability counts as an RTA for the models with fixed effects.

As the starting point to deviate from the probit model was the potential non-linearity to predict RTAs (see again motivational Figures 3 and 4), we now try to understand the captured non-linearity in the predictions of the NN.[19] Given the hidden layers and the complexity of the NN, this can not be done directly. We therefore use the predicted probabilities of 100 draws of 70 % samples of the training data and regress for each draw the predicted probabilities on RGDP_l, RGDP_s, POP_l, POP_s, DIST_{ls} up to the 20th polynomial as well as on DCONT_{ls}. Note that we did not use interaction terms and their polynomials as the computational effort required would grow to quickly, but the conclusion drawn from this exercise – namely that the process under consideration is highly non-linear – remains strong. Figure 15 plots the results in terms of the explanatory power of the included variables measured by the R^2. We see that up to the 10th polynomial the explanatory power increases and then by and large stabilizes at around 40 %. This implies that while we use all inputs of the NN as explanatory variables

[19] For the decision tree we already investigated the non-linearity by plotting the tree, see Figure 5.

and their polynomials to understand the predicted probabilities of the NN, we are only able to explain 40 % of the variation in the predicted probabilities of the NN. Hence, the NN captures besides the non-linearity that can be captured by polynomials of the explanatory variables also substantial complex interactions between predictors.

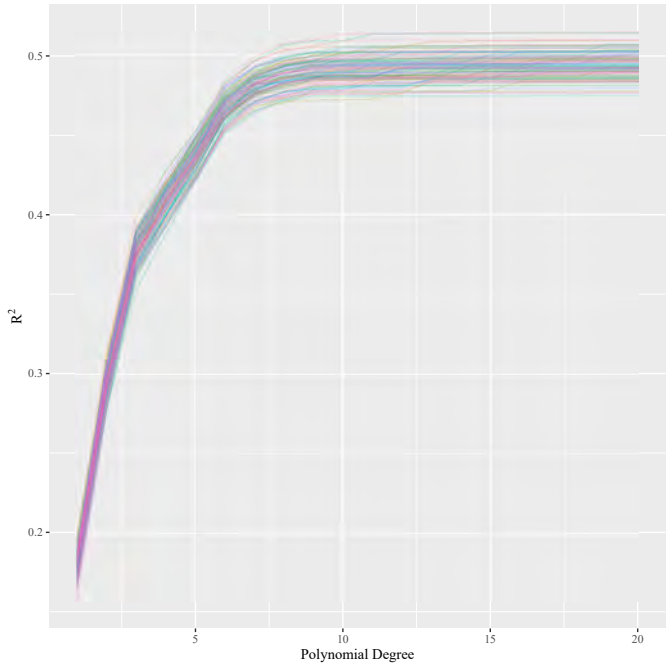

Figure 15: Goodness of fit of the predicted probabilities of the NN as a function of DCONT_{ls} and the polynomial degree up to order 20 of RGDP_l, RGDP_s, POP_l, POP_s, DIST_{ls} for 100 draws of 70% samples of the training data. Each draw is represented by one line.

Our last general prediction performance compares the predictions for different time periods. In training/estimating our models, we only used the year 2013. We now try to predict the probabilities to form an RTA for the different years, based on the values of the exogenous variables for the specific years. We use the models without fixed effects, as some countries are no longer in existence in 2013 and hence this would lead to missing fixed effects for these countries. Figure 16 plots, similarly as Figure 14, the true positives and true negatives for all three models for the years 1960 to 2018. The probit and NN predict nearly 100 % of the true positives in 1960, the gradient boosting machine a bit less (about 84 %). The reason is that a model trained on data for 2013 sees a lot more RTAs in place than was the case in the 1960s. Hence, it is more likely to predict RTAs, even with values of the exogenous variables from the 1960s. Interestingly, the share of true negatives is also comparably high, suggesting that the models do not work too badly even for early years. Over time, the share of true positives falls for all three models approximately until the 1990s, when they increase again. Hence, the models nicely capture the surge of RTAs in the 1990s, even though with some lag. The share of true negatives is for all models more stable over time. Of particular interest is the spike of the gradient boosting machine in 2013 which could be due to the aforementioned over-fitting of tree-based models on the training data. In terms of relative performance between the models, we see that there is no dominant model. However, both, the NN and the gradient boosting

machine, seem to be good in terms of both, true positives and true negatives likely due to its ability to generalize well in complex contexts. We view these results as quite encouraging, given that the models were trained with one year only. We believe that exploiting the time structure of the data also with machine learning methods is an exciting area for future research.

One concern when predicting on unseen data using machine learning methods such as NNs or gradient boosting is data leakage, where information that is not contained explicitly in the predictors is used by the algorithm to learn, providing the algorithm with information during training that is not usually available in other circumstances (see Kaufman et al., 2012). In our case this could occur if the methods in question identify certain countries using their economic heft and population. In particular one could construct a situation where a certain country negotiates relatively many RTAs. If said country is in the training as well as in the test data, the algorithm could attribute a higher likelihood of this country engaging in an RTA with another country not due to its size but solely due to the fact of it having many RTAs. Figure 16 provides evidence that this is not the case, as the predictors vary over time and hence are not identifying certain countries, yielding credibility to our reevaluation of RTA determinants derived from economic theory in previous chapters.

However, when introducing fixed effects, which we do for the prediction part of the paper, this issue remains unsolved, as the country identifier persists through time in the form of the country dummies. Yet, given that we are mostly interested in prediction and not theory evaluation in this section, one could argue that learning dyadic relationships from the dummies teaches the algorithm a form of network structure of global trade, provided that it increases prediction performance. One, and probably the largest, drawback is that the country identifier may not always be available, a problem to which we previously eluded to when discussing the construction of figure 16. For the purpose of this paper this is not much of a concern. We are however aware that this could lead to difficulties in other settings and should mostly serve to demonstrate the capabilities of modern machine learning methods to learn from high-dimensional data and most importantly to show that economic systems such as international trade are highly context dependent, making them a poor target for severe information compression. As can be seen in the immense improvement in prediction capabilities it is apparent that complex, idiosyncratic interactions between countries are also powerful drivers when it comes to conducting the negotiation of agreements. Furthermore, using methods that are able to cope with the breadth of information seems evident when trying to create a more accurate representation of the underlying data generating process.

Overall, we conclude that machine learning methods are quite capable to properly predict RTA formation. The trained machine learning models without fixed effects outperform our probit model without fixed effects in terms of true positive by 17.9 percentage points and in terms of true negatives by 18.6 percentage points. When introducing fixed effects, the NN predicts 41.6 percentage points more true positives and 0.9 percentage points more true negatives than the probit model.

4.2 Prediction Performances for Specific Events

Lastly, we use our three models for out of sample predictions for specific events. First, we use the European Union (EU) in its 2021 configuration as a bloc of 27 member countries and predict

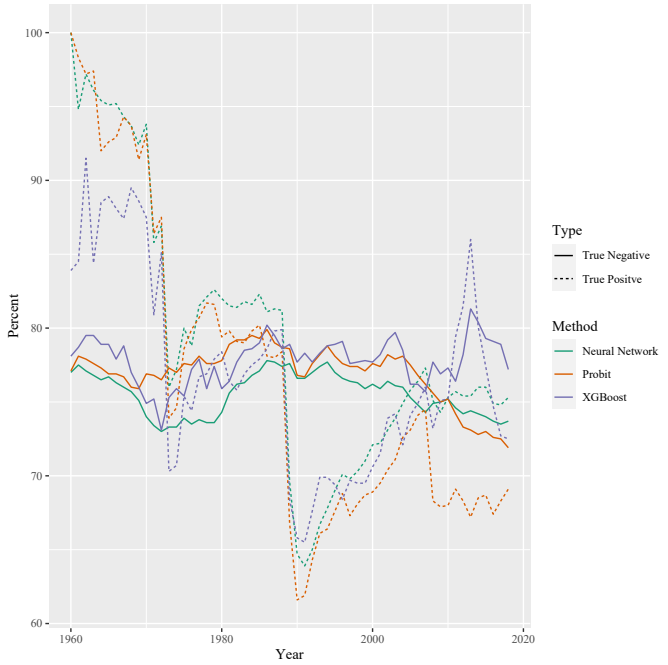

Figure 16: Out of sample predictions of all three models without fixed effects. The models were trained on data from 2013.

the likelihood of it concluding an RTA with other countries using our models with fixed effects. Our choice for predicting RTAs of the EU was motivated by the fact that the EU is very active in concluding new RTAs, which gives us the possibility to compare the prediction performance of our models in terms of currently negotiated agreements. In order to perform this analysis, we introduce a "voting mechanism" in which a member country votes for an RTA if its predicted probability for forming an RTA with the country under consideration is larger than 50 %. These votes are then cast by every member and aggregated. If the majority of EU countries is for an RTA we predict it as such. Relative to the fitting of the models, the novel aspects of this procedure are the use of 2018 data as well as the aggregation of the single EU member countries to one prediction.

Figures 17-19 illustrate the results of this exercise showing the implied probabilities that are calculated using the mean vote for a country by all EU countries. Given the graphic representation we conclude that the machine learning methods makes "bolder" decisions in that they produce estimates that are closer to either 0 or 1 than the probit. The gradient boosting machine is even more extreme than the NN in this respect. Hence, the machine learning methods give clearer predictions in terms of whether or not RTAs should be concluded according to the model. The country-specific predictions of the probit model also show that they are basically driven by the size of countries and somewhat by proximity, i.e., it mainly predicts RTAs with close and comparably large countries. Hence, the determinants of RTA formation that we found to be important are also reflected in the predictions. In the figure, where the country-specific predictions are aggregated using the mean vote, we see that probabilities are low for most countries. The NN and the gradient boosting machine seem to create clusters or regions with increased likelihood including Middle and South America, North Africa and the

Middle East as well as South East Asia. These countries seem to have similar, medium-sized GDP per capita, which could capture the increased trade relationships of these highly dynamic countries with all other countries in the world and specifically also with rich and large countries. Note that such trade relationships are consistent with trade theories based on factor endowment differences and productivity differences. Hence, also for the NN and gradient boosting we see that the predictions reflect the underlying determinants of RTA formation that we found in our previous analysis. The main visual differences from the graphs between the NN and the gradient boosting machine is that the gradient boosting machine is predicting higher probabilities across the board.

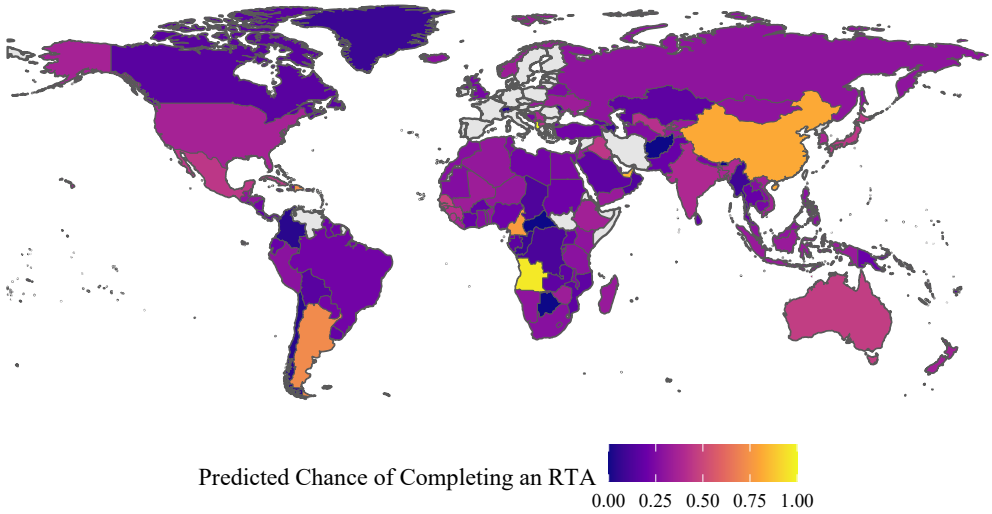

Figure 17: Prediction of the trained probit on the EU enacting an RTA given 2018 data.

Given these qualitative results, we use the confidence matrices in Table 7 for all three models for the exiting RTAs for the EU in the year 2018 in order to check how accurate there predictions are. While the probit and the gradient boosting machine do well in terms of true positives (87 % predicted correctly), the NN does a much better job in predicting actual non-RTAs correctly (97 %). Here, the probit model does a comparably bad job with only 14 % of the non-existing RTAs predicted as non-existing. Clearly, there is a trade-off between predicting zeroes and ones correctly. The NN seems to strike this balance best for predicting actual existing RTAs of the EU in the year 2018. This should be kept in mind when we next look at predictions of negotiated agreements.

Table 8 shows our models' predictions on EU recently concluded and ongoing trade negotiations as of March 2021.[20] The probit model predicts for only 1 out of 21 cases, Argentina, that the majority of countries would conclude an RTA.

The predictions of the NN show more variability. For 6 out of the 21 countries, Argentina, Chile, Mexico, Morocco, Tunisia and United Kingdom, the NN predicts an RTA. There are

[20] https://trade.ec.europa.eu/doclib/docs/2006/december/tradoc_118238.pdf.

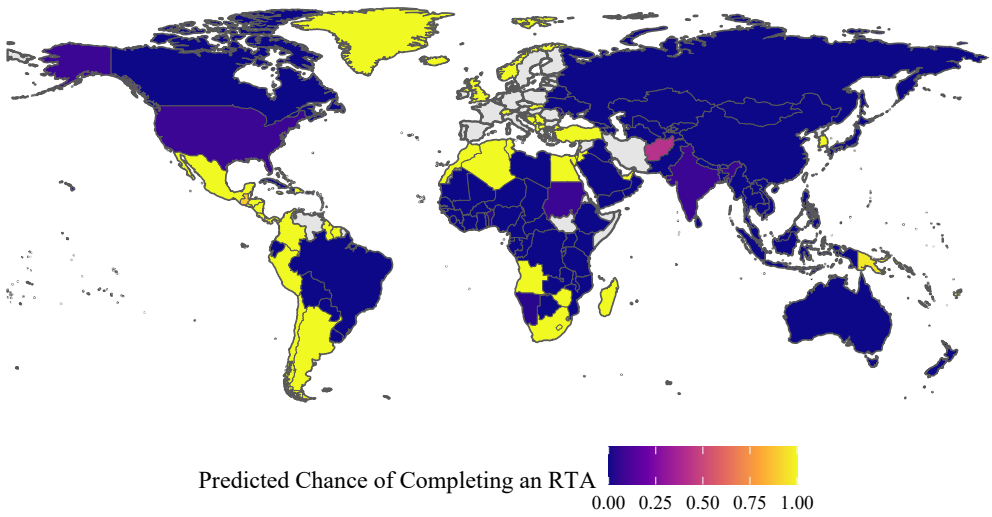

Figure 18: Prediction of the trained NN on the EU enacting an RTA given 2018 data.

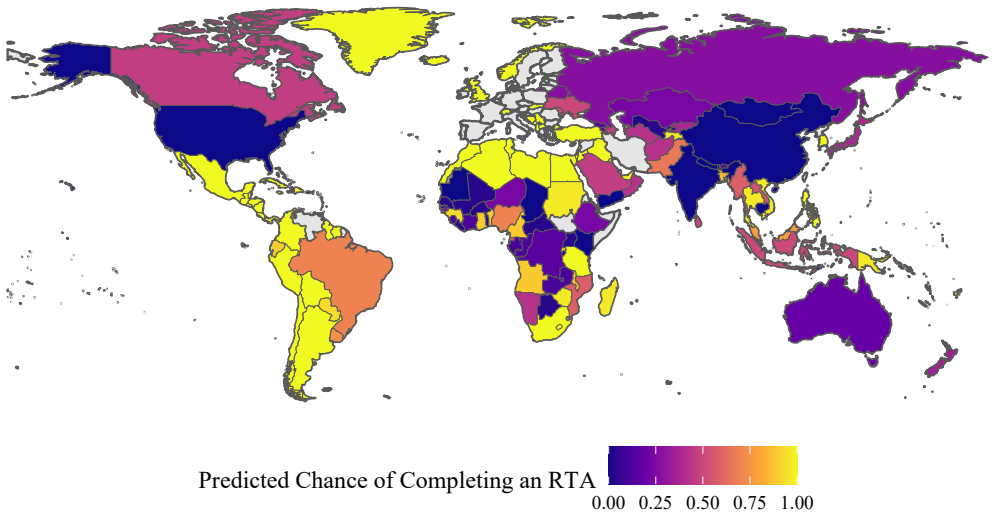

Figure 19: Prediction of the trained XGBoost on the EU enacting an RTA given 2018 data.

Table 7: Prediction performance on existing RTAs for the EU using the fixed effects models.

		Actual 0	Actual 1
Probit	Predicted 0	0.14	0.13
	Predicted 1	0.86	0.87
NN	Predicted 0	0.97	0.21
	Predicted 1	0.03	0.79
XGBoost	Predicted 0	0.67	0.13
	Predicted 1	0.33	0.87

several noteworthy observations concerning these findings: i) Since 1990 the EU has traded under the Framework Trade and Economic Co-operation Agreement with Argentina. Now, Argentina is part of the EU-MERCOSUR Association Agreement, for which an agreement in principle was reached in June 2019. Noteworthy, the EU is Argentina's third most important trading partner (after Brazil and China). ii) The negotiations with Chile are active and ongoing. The goal is to modernize the EU-Chile Association Agreement from 2017. iii) The negotiation of an agreement with Mexico is complete and will replace an already existing agreement (the EU-Mexico Global Agreement from 1997). Hence, there is already a tight relationship between Mexico and the EU, which the NN also seems to capture. iv) With Morocco the EU trades under the Association Agreement that entered into force in 2000. Now, since, 2013, negotiations for a Deep and Comprehensive Free Trade Area (DCFTA) take place. While the EU is Morocco's most important export market, for the EU Morocco ranks 22^{nd} as trading partner representing about 1,0 % of the EU's total trade with the world. v) In 2015, the EU and Tunisia launched negotiations for a Deep and Comprehensive Free Trade Area (DCFTA) building on the EU-Tunisia Association Agreement, which entered into force in 1998. Hence, there is already a tight relationship with Tunisia. vi) Similarly, the United Kingdom has of course a close relationship with the EU, and after its leave of the EU, the United Kingdom and the EU concluded the Trade and Cooperation Agreement in December 2020.

Comparing the results from the predictions of the NN and the gradient boosting machine, we see that the gradient boosting machine predicts nine RTAs more for the countries with whom the EU negotiates agreements as of 2019, i.e., for 15 out of the 21. Reassuringly, for all countries where the NN predicts an RTA, also the gradient boosting machine predicts and RTA. In addition, the gradient boosting machine predicts an RTA with Brazil, Malaysia, Myanmar, Paraguay, Philippines, Singapore, Thailand, Uruguay, and Vietnam. Brazil, Paraguay and Uruguay are all part of the of the EU-Mercosur Association Agreement. While Argentina and Brazil rank 40^{th} and 13^{th} in terms of total extra-EU trade, respectively, Paraguay and Uruguay rank only 109th and 72th, respectively.[21] For the Mercosur countries the EU is the number one trade and investment partner. The EU and the Mercosur states reached an agreement in June 2019. Malaysia, Myanmar, the Philippines, Singapore, Thailand, and Vietnam are all part of the ASEAN negotiating directives. ASEAN is one of the largest and most important free trade area in the world which also spurred further liberalization efforts, such as the Asia-Pacific Economic Cooperation, the East Asia Summit and the Regional Comprehensive Economic Partnership. Furthermore, the ASEAN trading bloc has numerous FTAs with other countries, such

[21] See https://trade.ec.europa.eu/doclib/docs/2006/september/tradoc_122530.pdf.

as Australia, China, India, Korea, and New Zealand.

For six of the 21 countries non of the models predicts an RTA: Australia, Canada, India, Indonesia, New Zealand, and USA. From this list, only the negotiations with Canada resulted in an agreement so far. The Comprehensive Economic and Trade Agreement with Canada is still not in force, as it is a mixed agreement and needs ratification from all member countries of the EU. The Transatlantic Trade and Investment Partnership with the USA did not come to light, and now there are two mandates to negotiate the elimination of tariffs for industrial goods and on conformity assessment. With Australia and New Zealand the negotiations are active and new rounds for negotiations have been set-up. The last round of negotiation with Indonesia were in June 2020. A date for a new round of negotiations is not yet agreed upon.

Table 8: Prediction of EU RTAs.

Country	Probit	NN	XGBoost
Argentina	1	1	1
Australia	0	0	0
Brazil	0	0	1
Canada	0	0	0
Chile	0	1	1
India	0	0	0
Indonesia	0	0	0
Malaysia	0	0	1
Mexico	0	1	1
Morocco	0	1	1
Myanmar	0	0	1
New Zealand	0	0	0
Paraguay	0	0	1
Philippines	0	0	1
Singapore	0	0	1
Thailand	0	0	1
Tunisia	0	1	1
UK	0	1	1
Uruguay	0	0	1
USA	0	0	0
Vietnam	0	0	1

Notes: Prediction form the models with fixed effects on whether the EU will enact an RTA with countries taken from the ongoing trade negotiations list. Data from https://trade.ec.europa.eu/doclib/docs/2006/december/tradoc_118238.pdf

Our second prediction performance for a specific event is predicting RTAs for the United States. The United States currently has 20 RTAs in force, 14 of which our fixed effects NN specification predicted with over 98 % certainty. It was participating in the negotiations of the Trans-Pacific Partnership (TPP), and is part of negotiations with the EU and a post-Brexit RTA with the United Kingdom, for example. Hence, it is also well suited as a specific example.

As we focus now on the RTA conclusions of a single country, the United States, and no longer a bloc of countries as in the case of the EU, we have the opportunity to directly compare

the predictions with and without fixed effects. Comparing Figures 20 and 21 we can draw several conclusions. Firstly, we see that Figure 20 seems to support the hypotheses promoted in a gravity trade context: distance seem to be prime determinants of RTA formation. Closer partners on the North American continent and countries easier to reach via sea in Europe and Asia are attributed higher likelihoods by our model. When turning to Figure 21 we notice that these effects seem somewhat muted. While close countries still get high probabilities, they are sometimes lower than from the model without fixed effects (Canada, for example). Other countries are now also obtaining quite high probabilities, even though they are comparably far away, such as India, China, and some European countries. Hence, the fixed effects capture some country-specific patterns that are only partly reflected in the NN without fixed effects. The predictions from the gradient boosting machine, Figure 22, also show high probabilities for neighboring countries and some South American countries. In addition, and in contrast to the NN, the gradient boosting machine predicts high probabilities for RTAs between the USA and South-East Asian countries as well as countries on the African continent. These are regions which both tend to have high numbers of RTAs. Another possible interpretation of Figure 22 is that the USA is predicted to form RTAs with nations that exhibit lower levels of GDP per capita, but not exceedingly so. This argument of complementary RTA formation would give reason to the improved prospect of China and India, both of which do not have a lot of RTAs (21 and 52, respectively), but could also help to answer the changes across Africa and Asia as a whole. It is thus possible to speculate that the gradient boosting machine has implicitly figured out some of the core theories of trade that were developed and have helped guide economic analysis for centuries. However, it is extremely likely that while the intuition behind such relationships reflected in the prediction probabilities seems very intuitive, the functional form is presumably very non-linear.

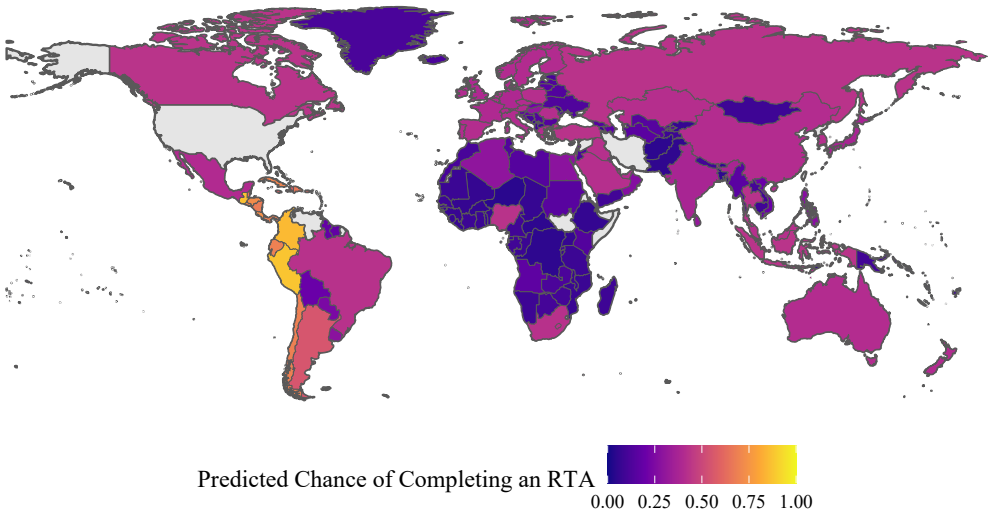

Figure 20: Prediction of the trained NN without fixed effects on the US enacting an RTA given 2018 data.

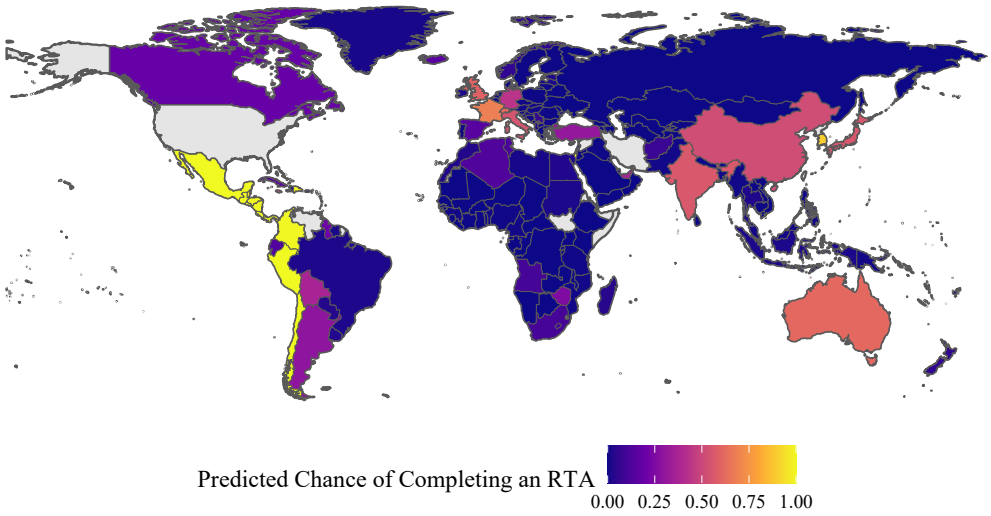

Figure 21: Prediction of the trained NN incorporating fixed effects on the USA enacting an RTA given 2018 data.

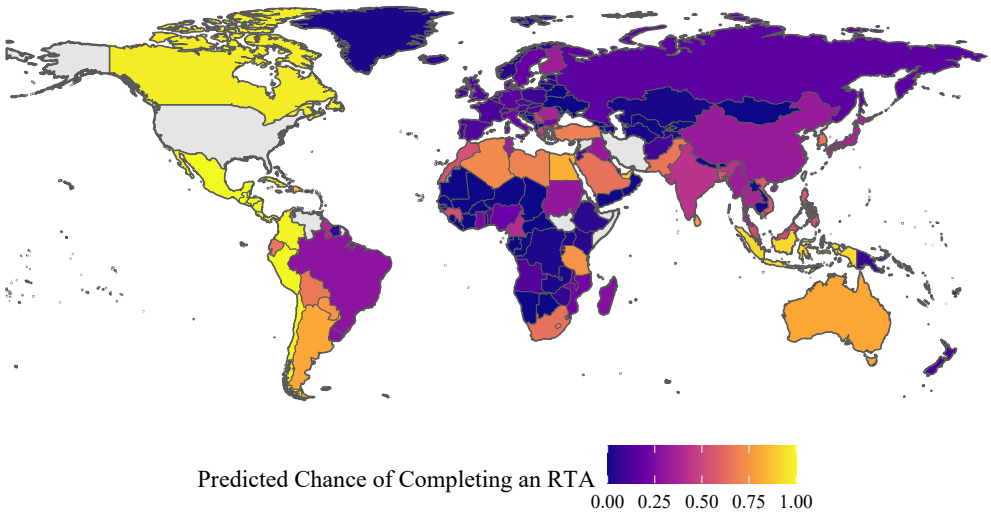

Figure 22: Prediction of the trained XGBoost incorporating fixed effects on the USA enacting an RTA given 2018 data.

As for the EU example Table, 9 shows the confidence matrix for all three models for the existing RTAs for the USA in the year 2018. All three models do comparably well in terms of true negatives, only the probit does poorly in terms of true positives. While the NN outperforms

the gradient boosting machine in predicting actual non-RTAs correctly (92 % versus 84 %), the gradient boosting machine does a better job in predicting actual RTAs (90 % versus 80 %). Once more we see the trade-off between predicting zeroes and ones correctly. For the US case, the gradient boosting machine does seem to strike this balance best, even though the performance of the NN and the gradient boosting machine are relatively similar.

Table 9: Prediction performance on existing RTAs for the US using the fixed effects models.

		Actual 0	Actual 1
Probit	Predicted 0	0.97	0.95
	Predicted 1	0.03	0.05
NN	Predicted 0	0.94	0.30
	Predicted 1	0.06	0.70
XGBoost	Predicted 0	0.84	0.10
	Predicted 1	0.15	0.90

Table 10 shows the predictions for the trade agreements that were in place in 2018. The cut-off value assigning ones is 0.5. From the 20 RTAs in place, the probit predicts only one, the RTA with Panama. The NN predicts 14 out of 20. Only for Bahrain, Jordan, Morocco, Oman, Singapore and Canada were the exiting RTAs not predicted. Three of these countries belong to the Middle East, and Morocco to Northern Africa. Hence, there maybe also some geopolitical reasons for the conclusion of these agreements which our NN did not capture. For some of these agreements, it is stated explicitly that the motivation is not only to spur trade but to "support [...] significant economic and political reforms"[22]. The gradient boosting machine also predicts ones for the 14 RTAs the NN predicts a one. In addition, the gradient boosting machine also predicts an RTA with Bahrain, Morocco, Singapore and Canada, which is in line with the observation that the NN seems to be more parsimonious than the gradient boosting machine. Yet, overall the dissimilar predictions of the probit and the machine learning methods emphasize the importance to take into account non-linearities when predicting RTAs. The strong overlap of predictions of the NN and the gradient boosting machine is encouraging, as it implies that different ways to allow for a more flexible modeling lead to similar outcomes. Overall, the good and similar prediction performances of the NN and gradient boosting machine discussed before are also reflected in the outcomes of predicting specific RTAs.

5 Conclusions

As has been made evident by the preceding analyses, machine learning techniques are capable of outperforming classical methods in predictive performance by a considerable margin thus being the superior description of the data generating process. We posit a multitude of arguments for this, which at their core refer to the ineptitude of classical approaches to map our highly complex world. Said relationships seem likely to follow functional forms that are beyond the understanding of any researcher. Yet, despite the limits that our study has highlighted, it also suggests that ideas and concepts at the core of economic thinking are of importance in determining RTA formation. The variables selected in traditional trade theory are able to

[22] https://ustr.gov/trade-agreements/free-trade-agreements/bahrain-fta, https://ustr.gov/trade-agreements/free-trade-agreements/morocco-fta.

Table 10: Prediction of existing United States RTAs.

Country	Probit	NN	XGBoost
Australia	0	1	1
Bahrain	0	0	1
Canada	0	0	1
Chile	0	1	1
Colombia	0	1	1
Costa Rica	0	1	1
Dominican Republic	0	1	1
El Salvador	0	1	1
Guatemala	0	1	1
Honduras	0	1	1
Israel	0	1	1
Jordan	0	0	0
Mexico	0	1	1
Morocco	0	0	1
Nicaragua	0	1	1
Oman	0	0	0
Panama	1	1	1
Peru	0	1	1
Singapore	0	0	1
South Korea	0	1	1

Notes: Using 2018 data to examine whether our models would have predicted existing United States RTAs using the fixed effects variants. Data from https://www.state.gov/trade-agreements/

explain a large amount of RTAs in our data, just not in the relationship as conventional analysis would have them. In particular we find support for theories of complementary trade between countries of dissimilar economic development. We conclude from this that many fundamental concepts of economic thinking are rightly reflected by empirical evidence, but may not be in line with the assumptions that are commonly posited in empirical analysis. Furthermore, we find that outcomes are highly context dependent, as our fixed effects approach lead to substantial different results and increases in predictive performance and hence likely representative power, both for classical and machine learning methods, supporting the argument of our economic system being governed by intricate relationships.

As already pointed out, the increase in performance comes at the cost of interpretability of the decision process, which we tried to illustrate in an intuitive manner. However, we propose that this is not a "bug" of the methods used herein but a feature of the world surrounding us and should not be regarded as an argument against such techniques but much more as a limit to our ability to understand such processes. Our analyses try to emphasize just how non-linear such a world can be. Modeling systems by means that are more easily to interpret at the cost of accuracy and more importantly truthfulness should not be the manner of economic inquiry if the ultimate purpose is to understand our surroundings.

Due to the reliable performance of approaches proposed here and other existing work we suggest to increase the use of machine learning tools in economic research contexts. One avenue of which could be to put to test further existing theories at the heart of economics to reveal their complex behavior.

References

Akman, E., A. S. Karaman, and C. Kuzey (2020): "Visa Trial of International Trade: Evidence from Support Vector Machines and Neural Networks," *Journal of Management Analytics*, 7, 231–252.

Alschner, W., J. Seiermann, and D. Skougarevskiy (2018): "Text of Trade Agreements (ToTA)-A Structured Corpus for the Text-as-Data Analysis of Preferential Trade Agreements," *Journal of Empirical Legal Studies*, 15, 648–666.

Athey, S. and G. Imbens (2017): "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31, 3–32.

Baier, S. and J. Bergstrand (2004): "The Economic Determinants of Free Trade Agreements," *Journal of International Economics*, 64, 29–63.

Baier, S. L., J. H. Bergstrand, and R. Mariutto (2014): "Economic Determinants of Free Trade Agreements Revisited: Distinguishing Sources of Interdependence," *Review of International Economics*, 22, 31–58.

Baldwin, R. and D. Jaimovich (2012): "Are Free Trade Agreements Contagious?" *Journal of International Economics*, 88, 1–16.

Batarseh, F., M. Gopinath, G. Nalluru, and J. Beckman (2019): "Application of Machine Learning in Forecasting International Trade Trends," *arXiv preprint arXiv:1910.03112*.

Bergé, L. (2018): "Efficient Estimation of Maximum Likelihood Models with Multiple Fixed-Effects: The R Package FENmlm," *CREA Discussion Paper Series available for download at https://EconPapers.repec.org/RePEc:luc:wpaper:18-13*.

Bernhofen, D. M., Z. El-Sahli, and R. Kneller (2016): "Estimating the Effects of the Container Revolution on World Trade," *Journal of International Economics*, 98, 36–50.

Cameron, A. and P. Trivedi (2005): *Microeconometrics - Methods and Applications*, Cambridge, United Kingdom: Cambridge University Press.

Chen, M. and S. Joshi (2010): "Third-Country Effects on the Formation of Free Trade Agreements," *Journal of International Economics*, 82, 238–248.

Chollet, F. and J. Allaire (2018): *Deep Learning with R*, Manning.

Circlaeys, S., C. Kanitkar, and D. Kumazawa (2017): "Bilateral Trade Flow Prediction," unpublished manuscript, available for download at *http://cs229.stanford.edu/proj2017/final-reports/5240224.pdf*.

De Veaux, R. D. and L. H. Ungar (1994): "Multicollinearity: A Tale of Two Nonparametric Regressions," in *Selecting Models from Data*, Springer, 393–402.

Efron, B. and T. Hastie (2016): *Computer Age Statistical Inference - Algorithms, Evidence, and Data Science*, Cambridge University Press.

Egger, P. and M. Larch (2008): "Interdependent Preferential Trade Agreement Memberships: An Empirical Analysis," *Journal of International Economics*, 76, 384–399.

Facchini, G., P. Silva, and G. Willmann (2013): "The Customs Union Issue: Why Do We Observe so Few of Them?" *Journal of International Economics*, 90, 136–147.

Goodfellow, I., J. Bengio, A. Courville, and F. Bach (2016): *Deep Learning*, MIT Press, Massachusetts.

Gopinath, M., F. Batarseh, and J. Beckman (2020): "Machine Learning in Gravity Models: An Application to Agricultural Trade," *NBER Working Paper No. 27151.*

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, New York: Springer, 2 ed.

Herbrich, R., M. Keilbach, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer (2001): "Neural Networks in Economics: Background, Applications and New Developments," *Department of Computer Science, Technical University of Berlin.*

Hummels, D. L. and G. Schaur (2013): "Time as a Trade Barrier," *American Economic Review*, 103, 2935–2959.

Jahn, M. (2018): "Artificial Neural Network Regression Models: Predicting GDP Growth," *HWWI Research Paper No. 185,* available for download at *https://www.hwwi.org/fileadmin/hwwi/Publikationen/Publikationen_PDFs_2018/.pdf.*

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An Introduction to Statistical Learning - with Applications in R*, New York: Springer, 1 ed.

Kaufman, S., S. Rosset, C. Perlich, and O. Stitelman (2012): "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6, 1–21.

Lahann, J., M. Scheid, and P. Fettke (2020): "Towards Optimal Free Trade Agreement Utilization through Deep Learning Techniques," in *Proceedings of the 53th Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences (HICSS-2020), January 7-10, Maui, Hawaii, United States*, ed. by T. Bui and R. Sprague, IEEE Computer Society.

Liu, X. (2008): "The Political Economy of Free Trade Agreements: An Empirical Investigation," *Journal of Economic Integration*, 23, 237–271.

Liu, X. and E. Ornelas (2014): "Free Trade Agreements and the Consolidation of Democracy," *American Economic Journal: Macroeconomics*, 6, 29–70.

Maggi, G. (2014): "International Trade Agreements," Chapter 6 in the Handbook of International Economics Vol. 4, eds. Gita Gopinath, Elhanan Helpman, and Kenneth S. Rogoff, Elsevier Ltd., Oxford, 317–390.

Maggi, G. and A. Rodríguez-Clare (2007): "A Political-Economy Theory of Trade Agreements," *The American Economic Review*, 97, 1374–1406.

Mullainathan, S. and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

Murphy, K. (2012): *Machine Learning - A Probabilistic Perspective*, MIT Press, Cambridge Massachusetts.

Quimba, F. and M. Barral (2018): "Exploring Neural Network Models in Understanding Bilateral Trade in APEC: A Review of History and Concepts," *PIDS Discussion Paper No. 2018-33*.

Scrucca, L. (2017): "On Some Extensions to GA Package: Hybrid Optimisation, Parallelisation and Islands Evolution," *The R Journal*, 9, 187–206.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): "Dropout: A Simple Way to Prevent Neural Networks From Overfitting," *The Journal of Machine Learning Research*, 15, 1929–1958.

Stammann, A. (2017): "Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-way Fixed Effects," *unpublished manuscript, available at https://arxiv.org/abs/1707.01815*.

Stock, J. and M. Watson (2017): "Twenty Years of Time Series Econometrics in Ten Pictures," *Journal of Economic Perspectives*, 31, 59–86.

Storm, H., T. Heckelei, K. Baylis, and K. Mittenzwei (2019): "Identifying Effects of Farm Subsidies on Structural Change Using Neural Networks," *Agricultural and Resource Economics, Discussion Paper 2019:1,* available for download at *http://ageconsearch.umn.edu/record/287343*.

Taddy, M. (2019): *Business Data Science*, McGraw-Hill Education Ltd.

Varian, H. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28, 3–28.

Wohl, I. and J. Kennedy (2018): "Neural Network Analysis of International Trade," *Office of Industries Working Paper ID-049,* available for download at *https://www.usitc.gov/publications/332/working_papers/neural_networks_and_international_trade_-_compiled_draft_06.pdf*.

Online Appendix

A Collineartiy of RGDPsim and RGDPsum with Exporter and Importer Fixed Effects

To show the perfect multi-collinearity of RGDPsim and RGDPsum with exporter and importer fixed effects in a cross-section, we first write down the definitions:

$$\begin{aligned}
\text{RGDPsum}_{ls} &= \log\left(\text{RGDP}_l + \text{RGDP}_s\right), \\
\text{RGDPsim}_{ls} &= \log\left(1 - \left(\frac{\text{RGDP}_l}{\text{RGDP}_l + \text{RGDP}_s}\right)^2 - \left(\frac{\text{RGDP}_s}{\text{RGDP}_l + \text{RGDP}_s}\right)^2\right).
\end{aligned}$$

Let us reformulate RGDPsim as follows:

$$\begin{aligned}
\text{RRGDPsim}_{ls} &= \log\left(1 - \left(\frac{\text{RGDP}_l}{\text{RGDP}_l + \text{RGDP}_s}\right)^2 - \left(\frac{\text{RGDP}_s}{\text{RGDP}_l + \text{RGDP}_s}\right)^2\right) \\
&= \log\left(\frac{\left(\text{RGDP}_l + \text{RGDP}_s\right)^2 - \text{RGDP}_l^2 - \text{RGDP}_s^2}{\left(\text{RGDP}_l + \text{RGDP}_s\right)^2}\right) \\
&= \log\left(\left(\text{RGDP}_l + \text{RGDP}_s\right)^2 - \text{RGDP}_l^2 - \text{RGDP}_s^2\right) - 2\log\left(\text{RGDP}_l + \text{RGDP}_s\right) \\
&= \log\left(\text{RGDP}_l^2 + \text{RGDP}_s^2 + 2\text{RGDP}_l\text{RGDP}_s - \text{RGDP}_l^2 - \text{RGDP}_s^2\right) \\
&\quad - 2\log\left(\text{RGDP}_l + \text{RGDP}_s\right) \\
&= \log\left(2\text{RGDP}_l\text{RGDP}_s\right) - 2\log\left(\text{RGDP}_l + \text{RGDP}_s\right) \\
&= \log\left(2\right) + \log\left(\text{RGDP}_l\right) + \log\left(\text{RGDP}_s\right) - 2\log\left(\text{RGDP}_l + \text{RGDP}_s\right).
\end{aligned}$$

In this last expression for RGDPsim, we have a constant ($\log(2)$), variables that are captured by the exporter and importer fixed effects ($\log\left(\text{RGDP}_l\right)$, $\log\left(\text{RGDP}_s\right)$), and one term that is given by $\log\left(\text{RGDP}_l + \text{RGDP}_s\right)$. However, this is exactly RGDPsum. Hence, RGDPsim and RGDPsum are perfectly collinear in the presence of exporter and importer fixed effects.

B Collinearity of DROWKL with Exporter and Importer Fixed Effects

We now show the perfect multi-collinearity of DROWKL with exporter and importer fixed effects in a cross-section. To do so, let us start with the definition of DROWKL:

$$
\begin{aligned}
\mathrm{DROWKL}_{ls} &= 0.5 \times \left| \log\left(\frac{\sum_{k\neq l} \mathrm{RGDP}_k}{\sum_{k\neq l} \mathrm{POP}_k} \right) - \log\left(\frac{\mathrm{RGDP}_l}{\mathrm{POP}_l} \right) \right| \\
&+ 0.5 \times \left| \log\left(\frac{\sum_{k\neq s} \mathrm{RGDP}_k}{\sum_{k\neq s} \mathrm{POP}_k} \right) - \log\left(\frac{\mathrm{RGDP}_s}{\mathrm{POP}_s} \right) \right| \\
&= 0.5 \times \left| \log\left(\sum_{k\neq l} \mathrm{RGDP}_k \right) - \log\left(\sum_{k\neq l} \mathrm{POP}_k \right) - \log\left(\mathrm{RGDP}_l\right) + \log\left(\mathrm{POP}_l\right) \right| \\
&+ 0.5 \times \left| \log\left(\sum_{k\neq s} \mathrm{RGDP}_k \right) - \log\left(\sum_{k\neq s} \mathrm{POP}_k \right) - \log\left(\mathrm{RGDP}_s\right) + \log\left(\mathrm{POP}_s\right) \right| \\
&= 0.5 \times \left| \log\left(\mathrm{RGDPTOT} - \mathrm{RGDP}_l\right) - \log\left(\mathrm{POPTOT} - \mathrm{POP}_l\right) - \log\left(\mathrm{RGDP}_l\right) + \log\left(\mathrm{POP}_l\right) \right| \\
&+ 0.5 \times \left| \log\left(\mathrm{RGDPTOT} - \mathrm{RGDP}_s\right) - \log\left(\mathrm{POPTOT} - \mathrm{POP}_s\right) - \log\left(\mathrm{RGDP}_s\right) + \log\left(\mathrm{POP}_s\right) \right|,
\end{aligned}
$$

where $\mathrm{RGDPTOT} = \sum_k \mathrm{RGDP}_k$, and $\mathrm{POPTOT} = \sum_k \mathrm{POP}_k$. The first line only varies over l, and the second only over s. As the two lines are additive, the parts can be explained by exporter and importer fixed effects. This explains the perfect multi-collinearity of DROWKL with exporter and importer fixed effects in a cross-section.

C Additional Figures and Tables

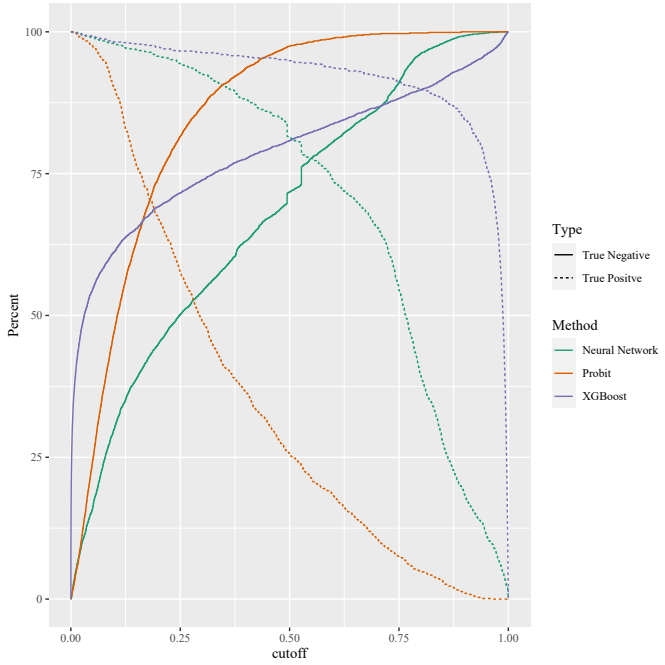

Figure 23: Trade off of true positives and negatives depending on the cut-off value at which the implied probability counts as an RTA for the models without fixed effects.

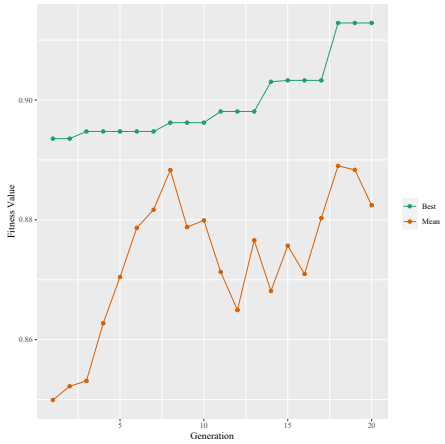

Figure 24: Evolution of the fitness function for the XGBoost using the AUC as a fitness metric.

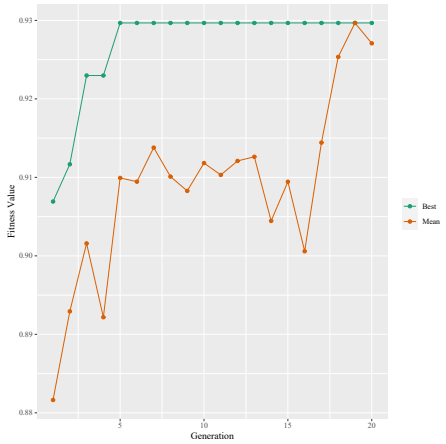

Figure 25: Evolution of the fitness function for the XGBoost with fixed effects using the AUC as a fitness metric.

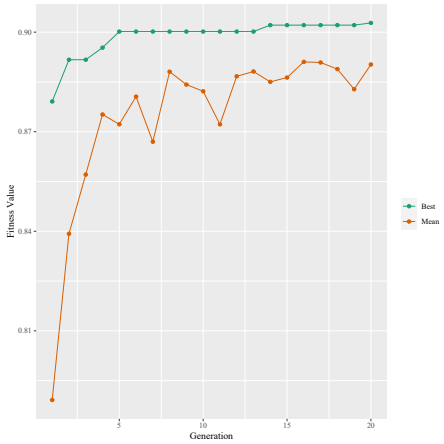

Figure 26: Evolution of the fitness function for the NN using the AUC as a fitness metric.

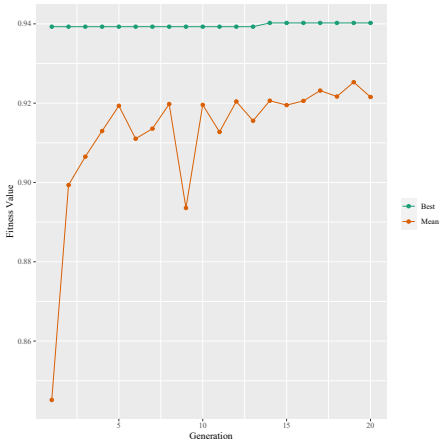

Figure 27: Evolution of the fitness function for the NN with fixed effects using the AUC as a fitness metric.

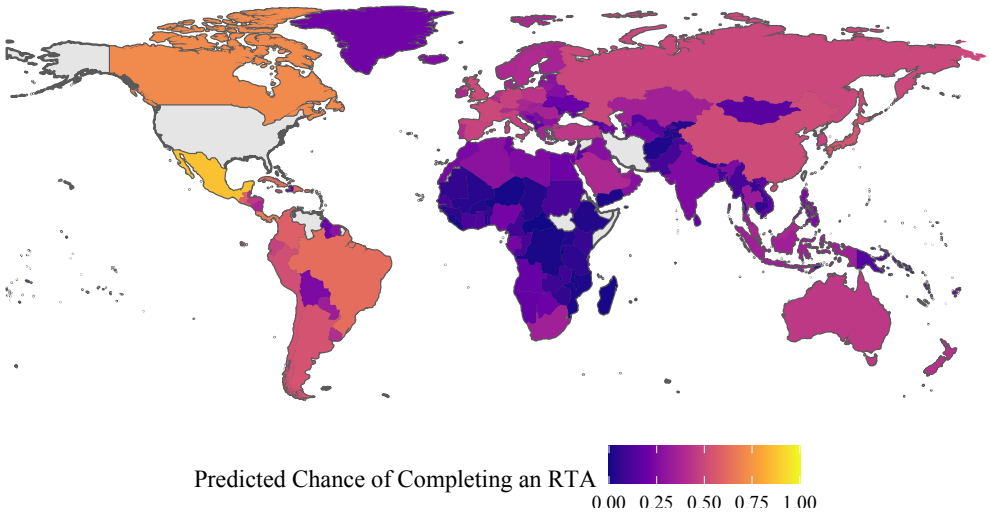

Figure 28: Prediction on US trade agreements using the probit without fixed effects.