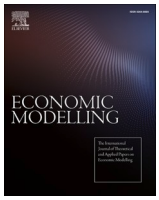
Economic Modelling 139 (2024) 106819

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling



Who will sign a double tax treaty next? A prediction based on economic determinants and machine learning algorithms☆

Dmitry Erokhin [a,b,*,1], Martin Zagler [a,c,1]

[a] WU Vienna University of Economics and Business, Austria
[b] International Institute for Applied Systems Analysis, Austria
[c] UPO University of Eastern Piedmont, Italy

ARTICLE INFO

Handling editor: Sushanta Mallick

Original content: Tax treaty formation (Original data)

JEL classification:
JEL: F53
H20

Keywords:
Machine learning
Treaty formation
Double tax treaty

ABSTRACT

Double tax treaties play a crucial role in shaping international economic relations, yet predicting which country pairs are likely to sign tax treaties remains a challenge. This study addresses this gap by employing a novel machine learning approach to predict tax treaty formations. Using data from a wide range of countries, we apply a series of classification algorithms and identify 59 country pairs likely to have tax treaties given their economic conditions. Our findings reveal that variables such as foreign direct investment, trade, Gross Domestic Product, and distance are significant predictors of tax treaty formations. Importantly, we demonstrate that the random forest classification algorithm outperforms conventional econometric methods in predicting tax treaty formations. By identifying which potential treaties exhibit a high probability of success, this paper gives policymakers an indication where to focus their attention and resources in upcoming treaty negotiations.

1. Introduction

Tax treaty formation is a very complex and multi-faceted decision-making process. Broadly speaking, the main goal of tax treaties is to boost trade and investment between countries by removing unnecessary tax barriers, which primarily means elimination of double taxation. Another important goal is to fight tax evasion and double non-taxation. In particular, new tax treaties focus more on anti-avoidance measures rather than on foreign direct investment (FDI) promotion (Blonigen and Davies, 2004). A third goal is an exchange of information, which is becoming the primary focus of new tax treaties and is also a subject of tax treaty negotiations. For illustration of different goals of a tax treaty, we can look at the explanation of a proposed treaty with Japan by the United States (Joint Committee on Taxation, 2004). That treaty defines goals of reduction or elimination of double taxation of income earned by residents of each country from sources within the other country, prevention of avoidance or evasion of the taxes of the two countries, promotion of closer economic cooperation between the two countries as well as elimination of possible barriers to trade and investment caused by overlapping taxing jurisdictions of the two countries. However, especially historically, tax treaty formation was also driven by "chess--games between superpowers", decisions of "key persons" (Evers, 2013), and corporate lobbyism (Thrall, 2021). Policy diffusion,[2] too, may have an effect on the policies adopted by the countries (Chen and Wang, 2021; Lopez-Cariboni and Cao, 2015) including in the area of taxation (Cao, 2010) and tax treaties (Barthel and Neumayer, 2012).

The significance of tax treaty formation and its implications for international economic relations cannot be overstated. As globalization continues to drive cross-border economic activities, the negotiation and formation of tax treaties between countries play a crucial role in

☆ This research is supported by the Austrian Science Fund (FWF): Doc 92-G. The authors are grateful for helpful comments and suggestions received at the Ghent Conference on International Taxation (Belgium), the XXIV Conference on International Economics (Spain), the Illinois International Accounting Symposium (Austria), the National Tax Association Conference (USA) as well as the DIBT research seminars at the Vienna University of Economics and Business. The authors acknowledge comments received from anonymous reviewers and editors.

* Corresponding author. WU Vienna University of Economics and Business, Austria.
E-mail addresses: erokhin@iiasa.ac.at (D. Erokhin), martin.zagler@gmail.com (M. Zagler).

[1] The authors have equally contributed to the paper.

[2] Policy diffusion means that countries follow their competitors or neighbors in terms of policies they adapt.

https://doi.org/10.1016/j.econmod.2024.106819
Received 27 July 2023; Received in revised form 25 April 2024; Accepted 3 July 2024
Available online 5 July 2024

facilitating international trade and investment, as well as in preventing double taxation and tax evasion. Understanding the dynamics of tax treaty formation is essential for policymakers, businesses, and investors seeking to navigate the complexities of international taxation and cross-border economic activities. To deal with this complex decision-making process, countries need to allocate a substantive number of resources to it. However, the capacity of treaty negotiators, especially in developing countries, is often limited. By employing a novel machine learning approach, this research aims to support decision makers and to shed light on the factors influencing the signing of tax treaties between countries and to provide predictive assessments of which country pairs are likely to have tax treaties. The findings of this study have the potential to inform policymakers and stakeholders about the patterns and determinants of tax treaty formation, thereby contributing to the development of more effective and informed international tax policies and economic strategies. First, this gives an indication to policymakers which treaties to pursue. Second, in the case of a capital importing income, if we find high probability that a neighbor will sign a DTT, there is a concrete risk that FDI will be diverted away from our economy to a neighboring jurisdiction. Third, in the case of a capital exporting economy, if a neighbor signs a DTT with a capital importing economy, our multinational firms will no longer find a level playing field in the foreign market. Understanding which countries are likely to sign a DTT in the future is crucial for economic policy.

To do it, the paper uses a novel method of machine learning. It applies the Stata/Python integration and implements a series of classification algorithms, in particular, classification tree, random forest, boosting, regularized multinomial, nearest neighbor, neural network, naive Bayes, support vector machine, and standard multinomial algorithms. The paper compares the algorithms in terms of their testing classification error rate and selects the random forest classification as the most accurate one. It then uses this algorithm to predict, which country pairs would have been likely to have a tax treaty in 2019. It identifies 59 country pairs likely to have tax treaties based on their features. Countries/regions with the highest number of predicted new tax treaties are Germany (9), Saudi Arabia (8), Brazil (7), Myanmar (7), and Hong Kong (6). In the discussion section, the results of the machine learning findings are discussed from the point of the current tax treaty status of the identified country pairs. Out of these identified country pairs, 31 are known to lead tax treaty negotiations, to have initialled a tax treaty, or to have already signed a tax treaty, 6 country pairs have signed or are negotiating an exchange of information agreement or a transport tax treaty, 3 country pairs used to have tax treaties, which were terminated. This supports the validity of the machine learning techniques for prediction purposes and makes them a relevant tool for policy makers.

Even though the focus of the paper lies on tax treaties, it is an illustration of how machine learning can be applied to support decision making as well as make policy predictions (Delogu et al., 2024; Zhang et al., 2023; Kleinberg et al., 2015). This makes the paper of a relevance and interest for a general audience not only those focused on international taxation.

The structure of the paper is as follows. Section 2 summarizes empirical literature on tax treaty formation. Section 3 describes data and machine learning approach. Section 4 presents the results. Section 5 makes predictions on which country pairs are likely to sign tax treaties in the future and discusses their policy relevance. Finally, Section 6 concludes.

2. Literature review

The formation of tax treaties has been the subject of empirical exploration in a limited but significant body of literature. Ligthart et al. (2011) conducted a pioneering study on the factors influencing countries' decisions to enter into tax treaties. Their extensive analysis, spanning 17766 country pairs from 1950 to 2006, revealed that the probability of countries signing tax treaties increases in response to various factors, including personal tax rates, non-resident withholding tax rates on dividends and interest, FDI stock, symmetric allocation of FDI, and a common language. This study laid the groundwork by highlighting the importance of economic and cultural ties in tax treaty formation.

Building on Ligthart et al. (2011) work, subsequent research expanded the understanding of tax treaty dynamics. Barthel and Neumayer (2012) analyzed 17205 country pairs between 1969 and 2005. Their study uncovered spatial spillovers in tax treaty formation, indicating that the likelihood of countries entering into tax treaties increased with the number of tax treaties signed by their regional peers and export-product competitors. Elsayyad (2012) introduced a bargaining model to analyze tax treaty formation between tax havens and Organisation for Economic Co-operation and Development (OECD) countries. Her research of 1323 country pairs identified tax haven bargaining power and good governance as the primary determinants of signing tax treaties.

Paolini et al. (2016) and Braun and Zagler (2018) then shifted focus towards the content of tax treaties, particularly the conditions under which they are signed, including information sharing and tax audits. Their research highlighted the delicate balance countries navigate between safeguarding revenue and facilitating international investment. In particular, Paolini et al. (2016) found that the likelihood of tax treaties between developing and developed countries increased with differences in tax rates between countries and decreased with transfer pricing, auditing costs, and average production costs. Braun and Zagler (2018) demonstrated that developed countries compensate developing countries for tax base losses resulting from tax treaties. Their study focused on 293 tax treaties signed between 19 donor and 68 recipient countries in the 1991–2012 period. Hearson (2018) contributed to our understanding of tax treaty formation by highlighting the impact of a government's revenue base, reliance on corporate tax, experience in signing tax treaties, and power asymmetries between signatories on the probability of signing a tax treaty and its content.

In addition to examining the formation of tax treaties, some studies delved into the specific content of these agreements, particularly the negotiated withholding tax rates. Studies by Petkova et al. (2020), Petkova (2021) and Chisik and Davies (2004) revealed how competition and investment asymmetry influence these rates, offering insights into the strategic considerations underpinning treaty negotiations. In particular, Petkova et al. (2020) analyzed withholding tax rates in over 3000 tax treaties and amending protocols between 1930 and 2012 and found a positive relationship with tax rates negotiated by competitors in previous tax treaties. Petkova (2021) identified spatial dependence in dividends withholding tax rates based on the tax rates of countries' peers. Chisik and Davies (2004) explored negotiated withholding tax rates and revealed that they increased as countries became more asymmetric in their foreign direct investment activities. Rixen and Schwarz (2009) reinforced this idea and showed a similar result for Germany's withholding tax rates with its 45 tax treaty partners signed up to 2003, where FDI asymmetries increased negotiated withholding tax rates.

Collectively, these studies, along with theoretical and legal discussions, underscore the complex decision-making process behind tax treaty formation. They have employed diverse methodological approaches to the topic, providing valuable insights into the factors influencing the signing of tax treaties. Our unique contribution to this literature is the application of a novel machine learning approach. We aim to identify the features of country pairs entering into tax treaties and make predictive assessments based on these features, shedding light on which country pairs are likely to sign tax treaties in the future. This approach extends the analytical toolkit and enhances our understanding of tax treaty formation dynamics.

3. Data and methodology

In the last few years, machine learning (ML) started gaining increased attention in the field of economics. Though there are older studies (e.g., Galindo and Tamayo, 2000 on credit risk assessment), economists remained cautious about the application of ML (Athey and Imbens, 2019). Now, it has already been applied in energy economics (e.g., prediction of crude oil and electricity prices, forecasting natural gas consumption) (Beyca et al., 2019; Ghoddusi et al., 2019), growth economics (e.g., forecast of US GDP growth, Japan GDP growth) (Soybilgen and Yazgan, 2020; Yoon, 2020), crypto economics (e.g., prediction of Bitcoin prices) (Chen et al., 2020), urban economics (e.g., analysis of historical data sources) (Combes et al., 2022), and many other areas of economics (Gogas and Papadimitriou, 2021). Machine learning techniques have also already found application in the area of taxation. Machine learning can be used to determine the optimal tax rate (Kasy, 2018), to predict tax crime and detect tax fraud, tax evasion and tax avoidance (Masrom et al., 2022; Zumaya et al., 2021; Ippolito and Lozano, 2020; De Roux et al., 2018), for tax planning and tax dispute resolution (Alarie and Xue Griffin, 2022; Alarie et al., 2016), to optimize tax administration policies (Battiston et al., 2024), to estimate effectiveness of taxation and tax reforms (Abrell et al., 2022; Lu et al., 2019; Andini et al., 2018; Zheng et al., 2016), to estimate the effect of taxes on prices and migration (Hull and Grodecka-Messi, 2022), to predict tax default (Abedin et al., 2022) and for many other purposes (Milner and Berg, 2017).

Given the complexity of the decision to enter into a tax treaty discussed above, machine learning seems to be a suitable mechanism to analyze country pairs with and without tax treaties against its possibility to model complex and more flexible relationships than simple linear models (Varian, 2014). Moreover, the primary goal of machine learning is prediction, which is in line with our intention to predict country pairs, which are likely to have a tax treaty based on their features. Whereas an economist would think first of a linear or logistic regression, non-linear machine learning techniques may actually be a better choice and allow uncovering generalizable patterns as well as finding functions that have a high out-of-sample predictive power (Mullainathan and Spiess, 2017). The out-of-sample predictability is of a high importance for policy makers who are in the first place interested in the effect of a policy on future outcomes and not so much in regression tables, which tend to neglect out-of-sample predictability (Basuchoudhary et al., 2017). It may well be the case that variables are highly significant but have a very poor out-of-sample fit, which questions the generalizability of the underlying model. In contrast to theory-driven deductive reasoning, machine learning lets the data speak (Cerulli, 2021a; Mullainathan and Spiess, 2017).

For example, the use of machine learning techniques in the prediction of economic growth demonstrates the benefits of machine learning techniques (Basuchoudhary et al., 2017). Given the variety of theoretical models of economic growth, the question arises on the many assumptions when selecting variables to explain economic growth as well as the assumptions on the variable distribution, whereas machine learning techniques neither require any prior theoretical assumptions nor any major assumptions on the variable distribution.[3] They require the choice of variables to train the algorithms, which is validated through the out-of-sample fit for the randomly chosen test sample, i.e., a test sample the algorithm has never seen before. Machine learning is of a special benefit when the actual relationship is unknown or complex. Researchers can uncover novel insights and patterns that may not have been apparent with conventional methods, without needing to motivate the inclusion of each particular variable or make predictions about their expected signs. This makes it attractive to be applied for the analysis of the multiplex decision to sign a tax treaty.

The question of whether country pairs have a tax treaty or not is a binary classification problem. We classify country pairs into "having tax treaties" and "not having tax treaties". We use the c_ml_stata_cv command (Stata/Python integration) for implementing machine learning classification algorithms (Cerulli, 2021b). The command makes use of the Python scikit-learn application programming interface (API) (Pedregosa et al., 2011). The command allows implementing the following classification algorithms: classification tree, random forest, boosting, regularized multinomial, nearest neighbor, neural network, naive Bayes, support vector machine, and standard multinomial (Scikit-learn, 2022). These algorithms are first trained to identify country pairs with tax treaties and country pairs without tax treaties based on different features and then validated in a test sample. Poulakias (2021) applies the command to predict occupational automation risk. Zhou and Li (2022) use a related r_ml_stata_cv regression command (Cerulli, 2021c) to forecast the COVID-19 vaccine uptake rate in the US.

To consider a large set of factors describing country pairs, which enter or do not enter into tax treaties, we use explanatory variables from the Centre d'Etudes Prospectives et d'Informations Internationales (CEPII) Gravity Database (Conte et al., 2022) and the International Monetary Fund (IMF) for FDI data (IMF, 2023). The dependent variable is a dummy variable, which is equal to two if country pairs have a tax treaty in a given year and one otherwise.[4] We use Tax Treaties Explorer to extract data on tax treaties (Hearson, 2021).[5] We divide our dataset into two periods to implement machine learning and answer our research question. We select 2018 as the year for training the machine to identify country pairs, which have tax treaties, and country pairs, which do not have tax treaties, and 2019 to test how well the training was. We also look at which country pairs would have had tax treaties in 2019 based on their features but had not had them yet. In total, after dropping missing values, we have about 2800 country pairs with tax treaties, and 6200 country pairs without tax treaties. Although it is naturally the case that there are more country pairs without tax treaties than country pairs with tax treaties, we consider our data set representative (30% vs. 70%).

We end up with the following variables: contiguity, simple distance between most populated cities, common official or primary language, common language spoken by at least 9% of the population, common colonizer post 1945, religious proximity index, colonial relationship post 1945, common legal origins before 1991, common legal origins after 1991, common legal origins change in 1991, colonial or dependency relationship ever, same colonizer ever, sum and absolute difference of population, sum and absolute difference of gross domestic product (GDP), sum and absolute difference of GDP per capita, General Agreement on Tariffs and Trade (GATT) membership, World Trade Organization (WTO) membership, European Union (EU) membership, presence of a regional trade agreement (RTA), sum and absolute difference of trade, sum and absolute difference of FDI, absolute difference in costs of business start-up procedures, absolute difference in number of start-up procedures to register a business, absolute difference in days

[3] We have to admit that there are certain assumptions for a set of machine learning algorithms, e.g., the assumption of independent-and-identically-distributed observations for the support vector machines or the assumption of independent observations for the logistic regression. In our case, where we look at tax treaties over time, a country-pair from year t is highly likely to be dependent on the same country-pair from another year. Hence, the independence assumption might not be met. However, it would only imply that these algorithms are less applicable for the given classification problem, and other algorithms without these assumptions may suit better.

[4] It is the recommendation of the command to recode the dummy variable from zero-one to one-two (Cerulli, 2020).

[5] The Explorer includes data originally published in the Treaties & Models collection on the International Bureau of Fiscal Documentation Tax Research Platform (IBFD), which has the fullest collection of tax treaties available. Missing tax treaties (esp., between developed countries) in the Explorer are checked manually in the IBFD.

required to start a business. For the variables to make sense for the machine learning algorithms and prediction, we construct all of them as bilateral variables. See Table A1 in Appendix for the variable description and data sources and Table A2 for summary statistics of the variables.

Table A3 summarizes the means of the above variables for country pairs with and without tax treaties. Country pairs with tax treaties have a significantly higher FDI, trade, GDP, GDP per capita, and population sum and difference than country pairs without tax treaties. This suggests that country pairs with tax treaties are larger in terms of the above variables but also more asymmetric than country pairs without tax treaties. Countries in country pairs with tax treaties have a significantly lower distance between them. They are significantly more likely to be contiguous, to have an RTA, to be WTO, and EU members. They have a significantly lower difference in entry costs, time, and procedures. They are significantly more likely to have been in a colonial or dependency relationship, and to have common legal origins change in 1991. They are less likely to be GATT members, and more likely to have a common language spoken by at least 9% of the population, but at a lower significance level. The differences in common official or primary language, common religion, common colonizer, and common legal origins before and after 1991 as well same colonizer between the two groups are not significant.

We put data into the machine learning algorithm to launch the meta-learning process, which consists out of three learning processes: learning over the tuning parameter, which is optimally selected to minimize the classification error rate[6] of the learner; learning over the algorithm $f(\cdot)$ to explore alternative algorithms with potentially higher predicting accuracy; and learning over new additional information when we put new data into the algorithm and reiterate the whole process (Cerulli, 2022). We use the classification error rate on the test data for the choice of the best-performing algorithm, i.e., proportion of misclassified country pairs in our case. It shows us how good the algorithm performs in the out-of-sample prediction. The classification error rate on the training set, on the contrary, could be misleading due to potential overfitting and should not be used for the algorithm selection.

Below we briefly explain the machine learning algorithms used in this paper, which are classification tree, random forest, boosting, regularized multinomial, nearest neighbor, neural network, naive Bayes, support vector machine, and standard multinomial algorithms. We use supervised machine learning methods because we can label the outcome for training and testing – country pairs with tax treaties and country pairs without tax treaties. We use all methods provided by the c_ml_stata_cv.

Classification tree learns simple decision rules from the data to create a predictive model. Classification tree has no requirements for data, such as their distribution or independence. Non-statistical requirements include the requirement that the entire training dataset is considered the root at the beginning, followed by the splitting of the data in a recursive manner. The number of leaves (maximum tree depth) is the tuning parameter, which has to be specified to run a classification tree. Fig. 1 illustrates an example of a classification tree, which is used to analyze loan eligibility.

Random forest is made up of a collection of classification trees with each tree being built from a sample drawn from the training set with replacement. Individual classification trees are then combined through averaging. Random forest has no distribution requirements and can handle multimodal and skewed data. For a random forest, we need to specify the maximum tree depth, the maximum number of splitting features, and the number of bootstrapped trees. Fig. 2 illustrates a random forest classifier (Khan et al., 2021).

Boosting solves the problem of constructing a strong learner – a learner that is well correlated with the true structure – from the set of

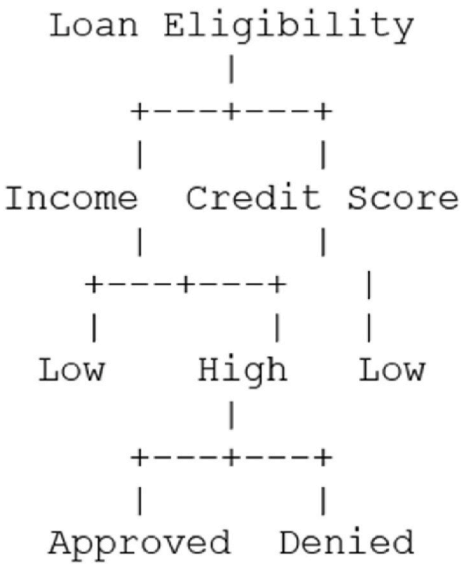


Fig. 1. Loan eligibility classification tree.

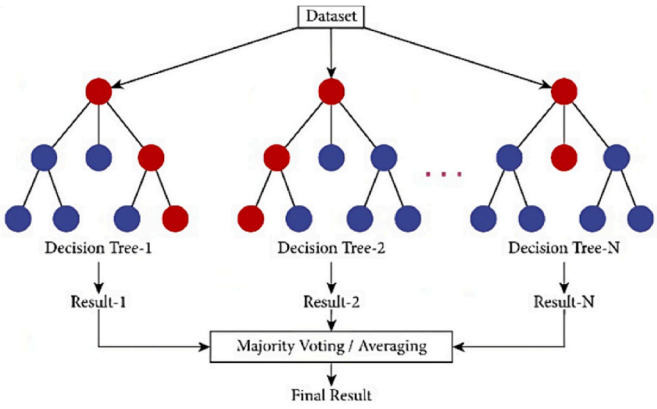


Fig. 2. Diagram of a random forest classifier (Khan et al., 2021).

weak learners – learners that perform only slightly better than random guessing (Schapire, 1990, 2003). In contrast to a random forest, boosting is a sequential algorithm (Scikit-learn, 2022). Boosting may assume an ordinal relationship between variable values. For boosting, we have to specify the maximum tree depth, the learning rate, and the number of sequential trees. Fig. 3 illustrates a boosting algorithm (Zhang et al., 2021).

Nearest neighbor is based on finding training samples closest to the new point and predicting its label based on these (Scikit-learn, 2022). Nearest neighbor assumes that data can be measured by distance metrics, and each training data point has a set of vectors and a class label. For nearest neighbor, we need to specify the number of nearest neighbors. Fig. 4 illustrates a nearest neighbor algorithm (Zhang, 2016).

Neural network is comprised of a set of nodes – neurons – and has an input layer, one or multiple hidden layers, and an output layer (Scikit-learn, 2022). Hidden layers are constructed from previous layers by a weighted summation of features. Neural networks do not have any assumptions on data. For neural network, we have to specify the number of neurons in the first layer, the number of neurons in the second layer, and the penalization parameter. Fig. 5 illustrates a neural network (Tanty and Desmukh, 2015).

Naïve Bayes is based on the application of the Bayes' theorem with the naive assumption of conditional independence between the features given the class variable (Scikit-learn, 2022). For example, an item may be considered a ball if it is round, white, and 22 cm in diameter. The

[6] *Classification error rate* $= \frac{false\ positives+false\ negatives}{true\ positives+true\ negatives+false\ positives+false\ negatives}$

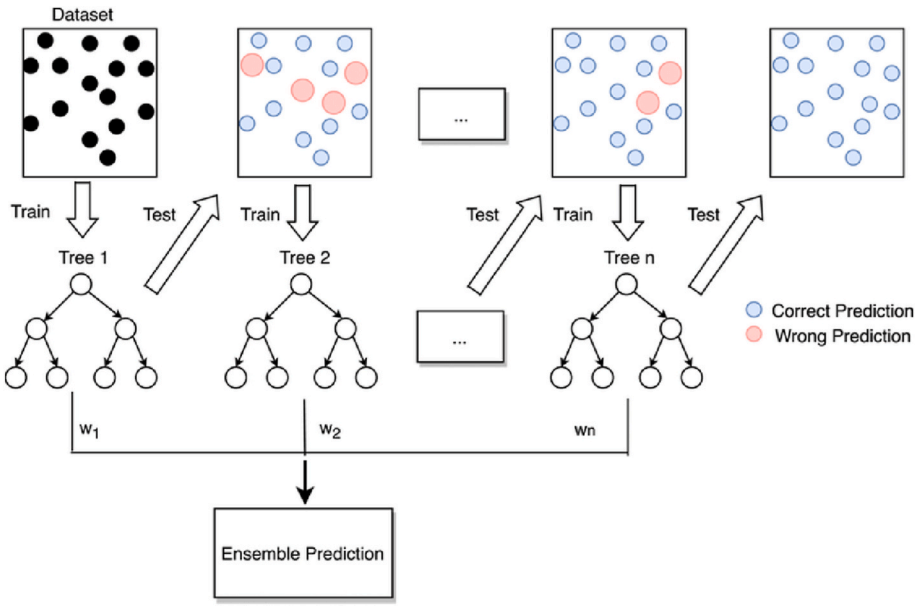


Fig. 3. Boosting algorithm (Zhang et al., 2021).

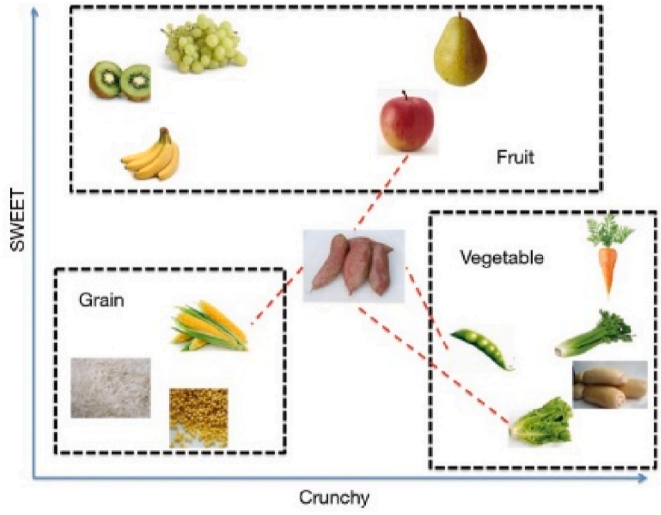


Fig. 4. Nearest neighbor algorithm (Zhang, 2016).

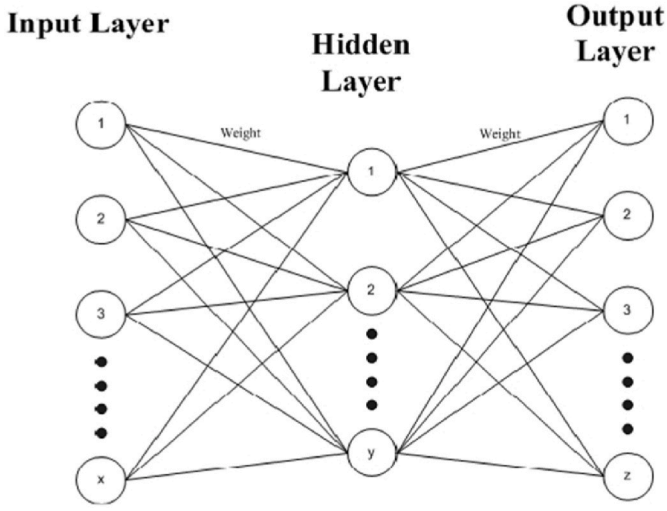


Fig. 5. Neural network (Tanty and Desmukh, 2015).

algorithm would treat all the three features (form, color, and diameter) separately to contribute to the probability of an item being a ball ignoring any possible correlations between the features. Fig. 6 illustrates a Naïve Bayes network in contrast with a Bayes network.

Standard multinomial performs a multinomial logistic classification (Scikit-learn, 2022). It is a general version of a binary logistic classification and is used to solve a classification task with multiple classes (two or more). Fig. 7 illustrates a standard multinomial algorithm. Inputs are transferred into logits using a linear model. The softmax function returns a probability that the observation belongs to the target class. Multinomial assumes that observations are independent and there is little or no multicollinearity among the variables.

Regularized multinomial is a version of the standard multinomial. The difference is that the algorithm is now regularized. Regularization penalizes model's complexity or smoothness and adjusts it in the way to reduce potential overfitting (Tian and Zhang, 2022; Bülhmann and van der Geer, 2011). For a regularized multinomial, we need to specify the penalization parameter, and the elastic parameter.

Support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space. The larger the distance to the nearest training data point, the better the separation between classes. Support vector machine assumes that data is independent and identically distributed. For support vector machine, we need to specify the margin parameter, and the inverse of the radius of influence of observations selected as support vectors. Fig. 8 illustrates the algorithm with three separating lines – support vectors (Scikit-learn, 2022). The solid line in the middle has the largest distance from both classes.

4. Main results of the machine learning algorithms

Table 1 summarizes the results of training and testing different machine learning algorithms in terms of their training and testing classification error rates. In total, we have 9057 training country pairs and 8787 testing country pairs. We run the above algorithms using default parameters.[7] We base the selection of the most accurate algorithm on the testing classification error rate because it shows how well an algorithm

[7] In the default mode, multinomial and regularized multinomial provide the same results. L2 regularization is applied by default.

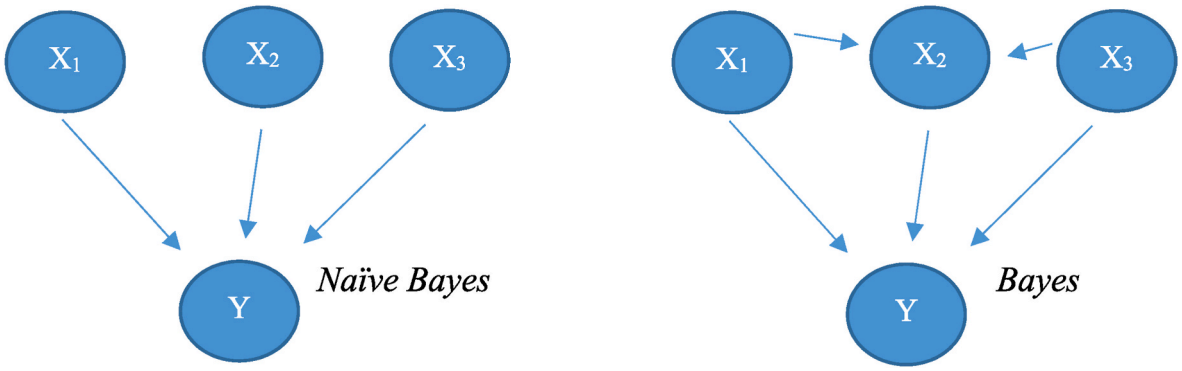


Fig. 6. Naïve Bayes network.

performs in the out-of-sample prediction. We see that it ranges between 0.057 and 0.273 with the random forest having the lowest testing classification error rate. Given the complex nature of the decision to enter into a tax treaty, it can be the case that there is no unique theory, which would predict tax treaty conclusion in every country pair. In such a case, the random forest performs best in the out-of-sample fit when variables may affect the outcome differently in different countries (Basuchoudhary et al., 2017).[8] Thus, we select the random forest algorithm for prediction. Given that our classes are imbalanced, in Table A5 (see Appendix), we summarize the results for alternative evaluation metrics such as sensitivity, precision, specificity, F1-score, and area-under-the-curve. The random forest algorithm outperforms other algorithms according to all the evaluation metrics.

Nonetheless, the suboptimal performance of certain algorithms can be attributed to the mismatch between their default settings and the characteristics of the data. To address this problem, we implement a hyperparameter selection process wherein we fine-tune the parameters of each algorithm with the goal of maximizing testing accuracy. This is achieved through the application of grid search in conjunction with a 10-fold cross-validation where an exhaustive set of possible features was tested (see Table 2). It is worth noting that random forest consistently outperforms the other algorithms. When we substitute the optimal random forest parameters into the train-test, we get the testing CER of approximately 0.057 as in the default setting[9]

Given that the random forest algorithm exhibits a zero training classification error rate and the lowest testing classification error rate, we present it in more detail. As discussed above, a random forest represents an average of a collection of random classification trees, so we present a specific classification tree first. In Fig. 9, we present a classification tree with 11 nodes drawn using the CART® Classification of the Minitab statistics package.

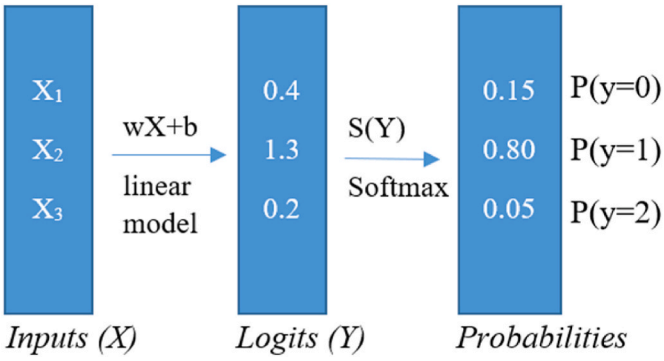


Fig. 7. Multinomial logistic classifier.

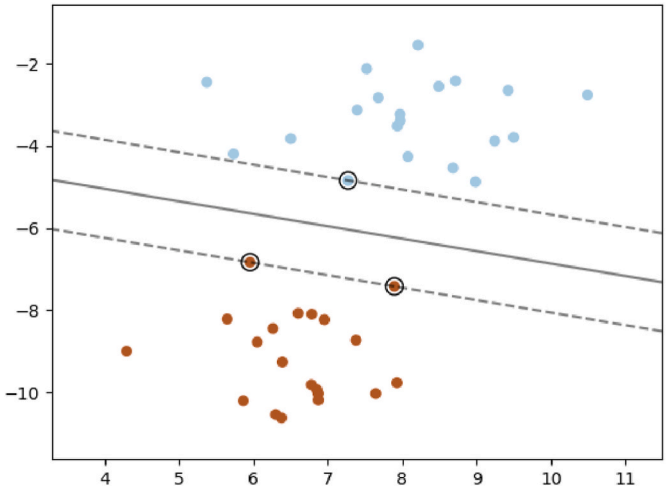


Fig. 8. Support vector machine algorithm (Scikit-learn, 2022).

In a classification tree, nodes are the key elements that make up the structure of the tree. There are two types of nodes: root nodes, which mark the beginning of the tree, and internal nodes (or decision nodes) that act as decision points. Each node contains information about a specific attribute or feature from the dataset and the rule for making decisions based on that attribute. Branches, on the other hand, are the connections that link nodes together. They represent the possible outcomes or categories that result from the decisions made at the parent node. When data is split at an internal node, branches connect to child nodes, which can be either other internal nodes or leaf nodes. The number of branches leaving an internal node depends on the number of possible outcomes for the attribute under consideration. In a classification tree, this hierarchical structure of nodes and branches is used to

[8] Though random forest outperforms other algorithms in this problem and also seems reasonable from a theoretical point of view, we cannot exclude the potential existence of other algorithms not covered in the paper, which could reach a higher accuracy.

[9] To be precise, the default testing CER is 0.0566, the testing CER for the model with optimal parameters is 0.0569.

[10] The prediction accuracy of an algorithm without tuning may be greater than of an algorithm with tuning due to a concept called overfitting. Overfitting occurs when a model is trained too well on the training data, and as a result, it performs poorly on new and unseen data. By tuning the parameters of the algorithm, the model may become more complex and may overfit the training data. In contrast, a model without tuning may be less complex and therefore less prone to overfitting.

[11] Another reason for the worse performance of some cross-validation in comparison to default models could be the limited variability in certain factors, e.g., distance. It could make it easier for the machine to classify countries in the whole sample. To address this limitation, we conduct an additional analysis where we only look at newly signed tax treaties.

Table 1
Predictive accuracy of machine learning methods.

Method	Number of used training units	Training classification error rate (CER)	Number of used testing units	Testing classification error rate (CER)
Random forest	9057	0	8787	0.057
Classification tree	9057	0	8787	0.108
Boosting	9057	0.144	8787	0.150
Nearest Neighbor	9057	0.175	8787	0.229
Regularized multinomial	9057	0.201	8787	0.200
Standard multinomial	9057	0.201	8787	0.200
Neural network	9057	0.245	8787	0.248
Support vector machine	9057	0.247	8787	0.250
Naive Bayes	9057	0.279	8787	0.273

Table 2
Hyperparameter selection outcome (10-fold cross-validation),[10,11]

Method	Parameter 1	Parameter 2	Parameter 3	Cross-validation training accuracy	Cross-validation testing accuracy
Random forest	Optimal tree depth = 50	Optimal n. of splitting features = 4	Optimal n. of trees = 100	1	0.859
Classification tree	Optimal tree depth = 5			0.855	0.833
Boosting	Optimal learning rate = 0.1	Optimal n. of trees = 100	Optimal tree depth = 5	0.933	0.856
Nearest Neighbor	Optimal n. of nearest neighbors = 20	Optimal kernel function = uniform		0.792	0.751
Regularized multinomial	Optimal penalization parameter = 0.001	Optimal elastic parameter = 0		0.473	0.473
Neural network	Optimal n. of neurons in layer 1 = 400	Optimal n. of neurons in layer 2 = 1000	Optimal L2 penalization = 0	0.687	0.687
Support vector machine	Optimal C parameter = 0.001	Optimal GAMMA parameter = 0.001		0.687	0.687

systematically divide and categorize data, eventually leading to leaf nodes where the final classification labels are assigned to the data points.

In the left branch of the tree where countries trade little we have only 13% of country pairs with tax treaties. In terminal node 1, we have country pairs, which trade little and have a low FDI difference. In this node, only 8% of country pairs have a tax treaty (e.g.,[12] Albania-Latvia, Bahrain-Yemen, Kyrgyz Republic-Moldova). In terminal node 2, we have country pairs, which trade little but have a higher FDI difference and are geographically close. In this node, the probability of a tax treaty increases up to 70% (e.g., Greece-Moldova, Armenia-Lebanon, Armenia-Cyprus). In terminal node 3, we have country pairs, which trade little, are geographically far away and have a medium FDI difference. The probability of them having a tax treaty is 20% (e.g., Ghana-Ireland, Solomon Islands-United Kingdom, Belarus-Hong Kong). If the FDI difference between these countries is very high (see terminal node 4), the probability increases to 65% (e.g., Luxembourg-Uruguay, Canada-Zambia, Malta-Mauritius).

Summarizing the left branch of the classification tree, countries that trade little tend not to have a double tax treaty, unless they are geographically close or they have large difference in FDI, which can be an indication that one partner in the treaty is an important capital exporter (FDI source country), whereas the other country is a capital importer (FDI destination country).

In the right branch of the tree where countries trade a lot, we have 60% of country pairs with tax treaties and 40% without, so no clear indication yet. The difference in entry costs, measured as the cost to start a business in the respective country, allows us to establish a 70%–30% distinction. If countries have a similar attitude to business, identified by similar entry costs, we are likely to see a treaty, whereas otherwise it is not very likely. Once again, FDI and geographical distance allows us to further distinguish country pairs by their probability to have a double tax treaty.

In terminal node 5, we have country pairs, which trade a lot, have a low difference in entry costs and FDI, and are at a medium geographical distance. For them, the probability of having a tax treaty is 60% (e.g., China-Croatia, Austria-Iceland, Malaysia-Slovak Republic). If they are very far away geographically, the probability falls to 22% (see terminal node 6) (e.g., Mexico-Ukraine, New Zealand-Norway, Australia-Israel). In terminal node 7, we have country pairs, which trade a lot, have a low difference in entry costs, and a high difference in FDI (e.g., China-Pakistan, Bosnia and Herzegovina-United Kingdom, Bangladesh-India). For them, the probability is 86%. In terminal node 8, we have country pairs, which trade a lot, have a high difference in entry costs, have a low sum of FDI, did not have a common colonizer, and are geographically close (e.g., Bahrain-Egypt, Egypt-Poland, Jordan-Qatar). For them, the probability is 42%. If they are geographically distant, the probability decreases to 16% (see terminal node 9) (e.g., Ecuador-Germany, Denmark-Kenya, Belgium-Nigeria). In terminal node 10, we have country pairs, which trade a lot, have a high difference in entry costs, have a low FDI sum, and had the same colonizer (e.g., Central African Republic-France, Nigeria-United Kingdom, Sudan-United Kingdom). Their probability of having a tax treaty is 93%. Finally, in terminal node 11, we have country pairs, which trade a lot, have a high difference in entry costs, and have a high sum of FDI (e.g., Kazakhstan-Tajikistan, Ukraine-United Arab Emirates, Myanmar-Singapore). The probability for them is 52%.

A random forest is a collection of classification trees, just like the one presented above. Each tree in the forest is distinct, but we can summarize the results of the random forest algorithm by counting how often a particular variable appears relevant in every tree. Fig. 10 illustrates the relative variable importance of all explanatory variables of the random forest, which measures mean decrease in impurity[13] within each tree with respect to the top predictor.[14] Trade sum is identified as the most

[12] For illustration, three examples of country pairs were randomly selected for each terminal node from our data.

[13] The impurity measures how well a split in each variable divides the data into correct classes (Disha and Waheed, 2022).

[14] We use Stata pylearn module (pyforest function) to calculate the variable importance (Droste, 2020).

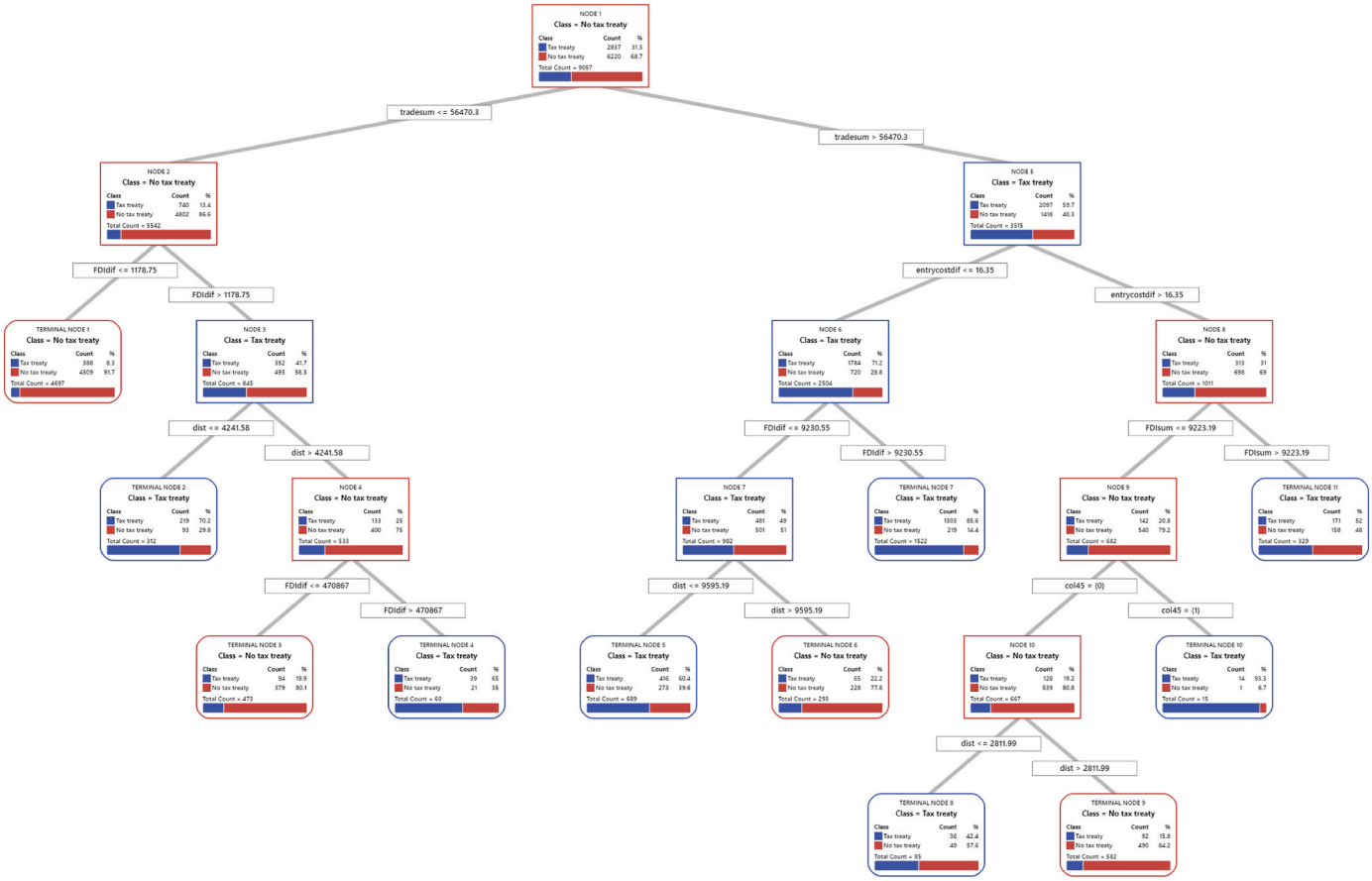


Fig. 9. Example of a classification tree with 11 nodes.

important variable (100.00% relative importance) followed by FDI difference (77.68%), distance (58.48%), and FDI sum (48.50%). Trade difference (45.91%) as well as entry cost difference (43.50%) and GDP per capita sum (30.23%) are also important. Under top ten variables, we also have GDP sum (28.97%), common religion (26.22%), and GDP difference (24.61%).

5. Prediction and policy implications

We can use the random forest model to calculate the probability that two countries should have a double tax treaty in place. And we can confront this with the actual data. Fig. 11 illustrates the share of country pairs with tax treaties in each predicted probability decile. We see that the share clearly increases with the probability, demonstrating again the validity of the algorithm. We can distinguish three groups. If the predicted probability is above 60%, than more than nine out of ten countries will actually have a double tax treaty in place. If the predicted probability to have a DTT is below 30%, than less than one out of ten countries will actually have a double tax treaty. Only if the probability to have a tax treaty is between 30% and 60%, we do see both countries pairs with and without treaties, as expected. Within this range, with increasing probability countries actually are more likely to have a treaty already in place.

Fig. 12 illustrates boxplots for country pairs with a tax and for country pairs without a tax treaty in 2019. We see that on average country pairs with a tax treaty have a much higher average predicted probability of having a tax treaty than countries without a tax treaty. The difference is statistically significant. This demonstrates the algorithm is very good at predicting the status of a particular country.

We can look beyond the sample horizon, which ended in 2019 due to data availability and see whether our results are an indicator of the likeliness of negotiations of double tax treaties. In Fig. 13 we draw boxplots for four different types of negotiation status. We present probabilities for treaties that have been signed since 2019 on the left, next treaties where negotiations have been completed (initialed), but the respective parliaments have not yet ratified them, then country pairs that are currently negotiating a treaty, and finally to the right country pairs that have not initiated treaty negotiations.

There is little difference between the first three categories. The medians are very similar, as much of distribution. Clearly, the start of negotiations can be predicted with our model, but not there completion. That depends much more on politics and resources devoted to negotiations, and even to timing. Note that the observation period falls into the global pandemic (2020–2022), and countries and negotiation teams had other priorities than to fly across the world to negotiate a double tax treaty. The last category stands out, with a much lower median with respect to the other three. Countries that do not negotiate a treaty clearly have a much lower incentive to do so. We present simple t-statistics between all four categories to test the null hypothesis whether the median of one category is statistically different from any other category (see Table 3). As we already saw in the boxplots, this is the case only for the last category (no treaty negotiations), at the 1% significance level.

Fig. 12 contains a few outliers, countries with a high probability to have a treaty that do not have one, and vice versa. Whilst the latter can be attributed to politics (for instance colonies), we can take a closer look at countries without treaties that exhibit a high probability to have one in place. We have already represented these cases in Fig. 13 in our boxplots. Table A6 in the appendix goes one step further and lists 59 country pairs that based on their 2019 features had been likely to have tax treaties (probability >60%) but had not had them yet. The prediction comes from the random forest. Column 3 contains the probability of the country pairs having a tax treaty in 2019, which ranges between

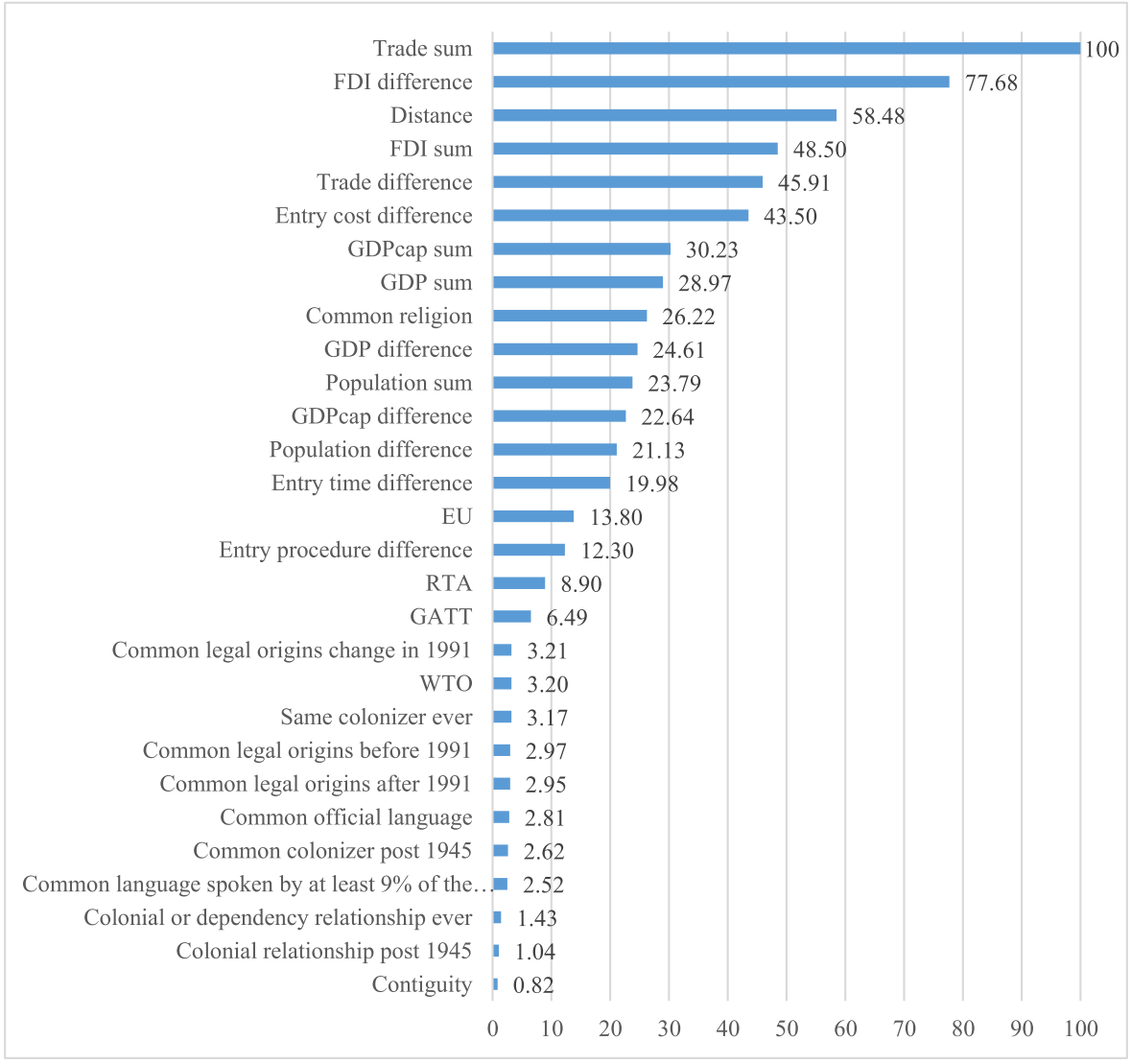


Fig. 10. Relative variable importance.

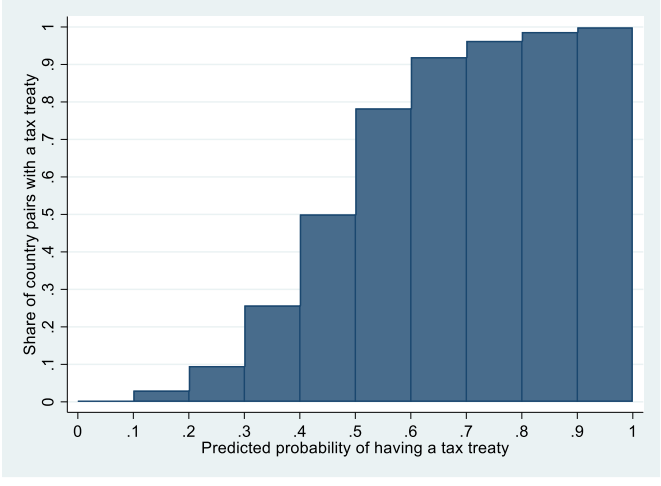


Fig. 11. Share of country pairs with tax treaties in each predicted probability decile.

Table 3
T-statistics.

Median/ Distribution	Signed	Initialed	Under negotiation	No tax treaty
Signed	.	0.2214	1.5750	4.4240***
Initialed	−0.2214	.	0.8421	3.6995***
Under negotiation	−1.5750	−0.8421	.	11.7288***
No tax treaty	−4.4240***	−3.6995***	−11.7288***	.

***Note*:** t-statistics. *** identifies significance at the 1% level.

0.60 and 0.93. The table is extended by the current tax treaty status of the country pairs in column 4. We see that 24 country pairs are in the negotiation process, 4 have signed a tax treaty, and 3 have initialed a tax treaty. 6 country pairs have signed or are negotiating an exchange of information agreement or a transport tax treaty. 3 country pairs used to have tax treaties, which were terminated. Especially appealing is the identification of the 19 country pairs, which are likely to conclude tax treaties in the future, but no negation has been reported at the date of this publication. This is particularly relevant for policymakers. First, it gives clear guidelines to the countries involved to understand which future treaty may pose an attractive opportunity. Second, it gives

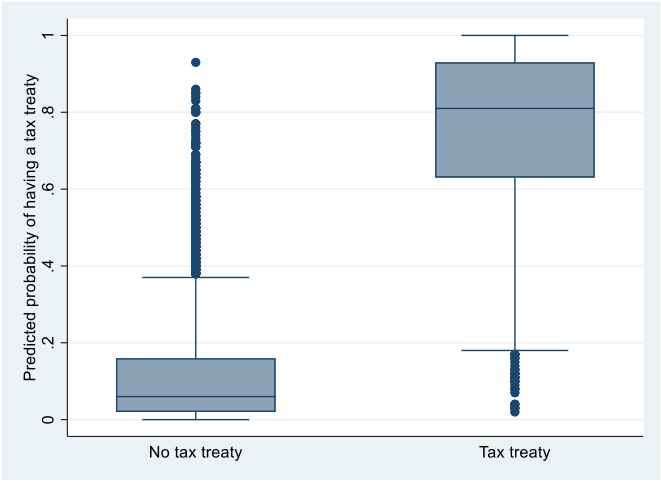


Fig. 12. Probability of having a tax treaty by actual tax treaty status.

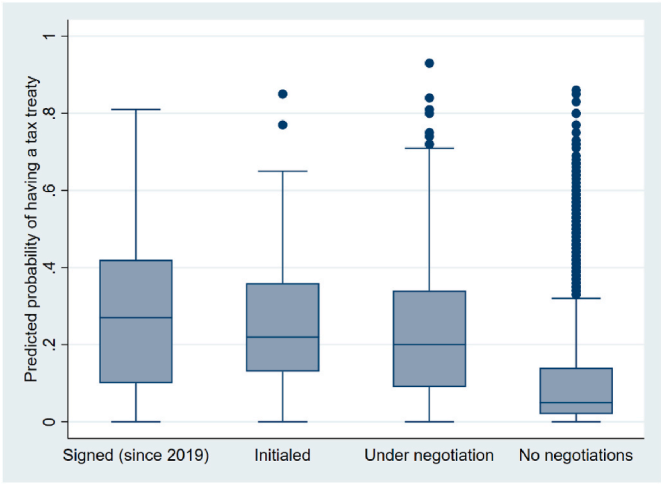


Fig. 13. Probability of having a tax treaty and current negotiation status.

neighboring countries an indication which potential treaty may be a competitive pressure for their respective economies.

The following countries/regions stand out by the number of predicted tax treaties: Germany (9), Saudi Arabia (8), Brazil (7), Myanmar (7), and Hong Kong (6). Below, we discuss the countries/regions with the highest number of predicted tax treaties.

Brazil has a population of over 214 million people and is a resource-abundant country. Brazil is under top 25 most attractive countries for FDI worldwide and the third one under the emerging markets (Kearney, 2022). However, the country only has 36 ratified tax treaties, though its number of tax treaties is increasing (Dagnese, 2006). For example, in the United Kingdom the lack of a tax treaty with Brazil was regarded as a gap in the UK global tax treaty network and its conclusion was seen as one of the main priorities (KPMG, 2022).

Germany as the largest country in the European Union both by population and GDP is clearly an attractive economic partner to have a tax treaty with. Moreover, Germany is regarded as the second most attractive destination for FDI globally (Kearney, 2022).

Hong Kong SAR (Special Administrative Region of the People's Republic of China) is a highly developed place with a network of 45 tax treaties. Hong Kong plays a dominant role of an intermediary for FDI flows in Asia (Leung & & Unteroberdoerster, 2008) and is one of the 8 major "pass-through economies" globally (Damgaard et al., 2018). It is both known as an FDI tax haven or offshore financial center (Hines Jr, 2010) and a place for round-tripping FDI (Xiao, 2004). Especially, it is an attractive conduit location to enter the Mainland China (Hong, 2018). The benefits and opportunities Hong Kong provides to foreign investors make it an attractive tax treaty partner.

Myanmar is a country with a population of 54 million people rich in resources like precious stones, rare-earth metals, oil, and natural gas.[15]

Table 4
Logit and probit regression coefficients.

Variables	(1)	(2)
	Logit	Probit
Logged trade sum	0.388***	0.214***
	(0.0259)	(0.0143)
Logged FDI difference	0.102***	0.0594***
	(0.00738)	(0.00424)
Distance	−0.000153***	−8.62e-05***
	(9.49e-06)	(5.25e-06)
Logged FDI sum	0.0537***	0.0307***
	(0.00769)	(0.00443)
Logged trade difference	−0.0704***	−0.0403***
	(0.0230)	(0.0129)
Entry cost difference	−0.0204***	−0.00979***
	(0.00188)	(0.000909)
Logged GDPcap sum	0.795***	0.453***
	(0.0896)	(0.0500)
Logged GDP sum	−0.546***	−0.313***
	(0.0639)	(0.0361)
Common religion	−0.128	−0.0660
	(0.137)	(0.0760)
Logged GDP difference	0.0433	0.0339
	(0.0396)	(0.0225)
Logged population sum	0.471***	0.260***
	(0.0824)	(0.0463)
Logged GDPcap difference	−0.0901**	−0.0538**
	(0.0389)	(0.0217)
Logged population difference	−0.0428	−0.0244
	(0.0387)	(0.0219)
Entry time difference	−0.00298*	−0.00137
	(0.00174)	(0.000948)
EU	0.105	0.0513
	(0.0735)	(0.0415)
Entry procedure difference	−0.0329**	−0.0194**
	(0.0145)	(0.00809)
RTA	0.252***	0.157***
	(0.0773)	(0.0438)
GATT	−0.293***	−0.156***
	(0.0654)	(0.0369)
Common legal origins change in 1991	0.825***	0.452***
	(0.134)	(0.0760)
WTO	−0.103	−0.0740
	(0.0965)	(0.0545)
Same colonizer ever	−0.0911	−0.0524
	(0.131)	(0.0729)
Common legal origins before 1991	0.0942	0.0619
	(0.125)	(0.0708)
Common legal origins after 1991	−0.333***	−0.178**
	(0.125)	(0.0707)
Common official language	0.222	0.124
	(0.168)	(0.0904)
Common colonizer post 1945	0.707***	0.397***
	(0.151)	(0.0836)
Common language spoken by at least 9% of the population	−0.284*	−0.163*
	(0.161)	(0.0872)
Colonial or dependency relationship ever	0.986**	0.680***
	(0.386)	(0.208)
Colonial relationship post 1945	1.504***	0.607**
	(0.470)	(0.241)
Contiguity	−0.484**	−0.310**
	(0.219)	(0.121)

[15] For example, Myanmar supplies up to 90% of world rubies (Shor and Weldon, 2009) and produces around 9% of the world's rare earths, which makes it the third largest rare-earth producer worldwide (US Geological Survey, 2022).

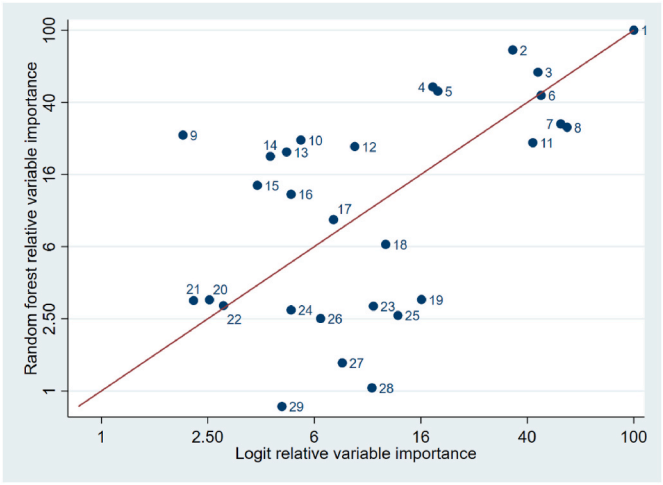


Fig. 14. Relative variable importance for random forest and logit regression. ***Legend***: 1. Trade sum, 2. FDI difference, 3. Distance, 4. FDI sum, 5. Trade difference, 6. Entry cost difference, 7. GDPcap sum, 8. GDP sum, 9. Common religion, 10. GDP difference, 11. Population sum, 12. GDPcap difference, 13. Population difference, 14. Entry time difference, 15. EU, 16. Entry procedure difference, 17. RTA, 18. GATT, 19. Common legal origins change in 1991, 20. WTO, 21. Same colonizer ever, 22. Common legal origins before 1991, 23. Common legal origins after 1991, 24. Common official language, 25. Common colonizer post 1945, 26. Common language spoken by at least 9% of the population, 27. Colonial or dependency relationship ever, 28. Colonial relationship post 1945, 29. Contiguity.

After becoming independent from the UK Myanmar experienced turbulent time with lasting civil-war periods. The liberalization of the country in the latest years led to weakening of Western sanctions and opening the country to the world. This would make it attractive for Myanmar to develop and deepen economic relations with other countries. However, the military coup in 2021 might postpone these developments.

Saudi Arabia is a resource-abound country with around 36 million people. For a long period of time, it used to have only one tax treaty with France (Daman, 2006). However, it has started expanding its tax treaty network to improve economic relations and attract more FDI.

Table 5
Predictive accuracy of machine learning methods.

Method	Number of used training units	Training CER	Number of used testing units	Testing CER
Random forest	671	0	5991	0.009
Classification tree	671	0	5991	0.082
Boosting	671	0.033	5991	0.035
Nearest Neighbor	Does not work with default parameters			
Regularized multinomial	Same as standard multinomial with default parameters			
Standard multinomial	671	0.083	5991	0.029
Neural network	671	0.206	5991	0.159
Support vector machine	671	0.073	5991	0.006
Naive Bayes	671	0.095	5991	0.060

6. Robustness check and conventional econometric methods

In addition to machine learning algorithms, we want to utilize traditional econometric methods in the analysis of tax treaty formation. In line with established literature, we employ logit and probit regressions to gain insights into the data. Whereas conventional econometric methods investigate global maximum likelihood, machine learning algorithms tend to search for local maxima, and therefore differ fundamentally in method. We could tackle this issue with a full set of interaction effects in a conventional econometric model estimation. With many variables, this would lead to a dramatic decline in the degrees of freedom, and hence technically infeasible. We therefore do not include any interaction effects here. Table 4 summarizes the regression coefficients for logit and probit.

We take logarithms of FDI, trade, GDP, GDP per capita, and population sum and difference, which would be a classical way when using these variables in a regression. We do not expect this approach to significantly impact our results. In addition, we use the estat

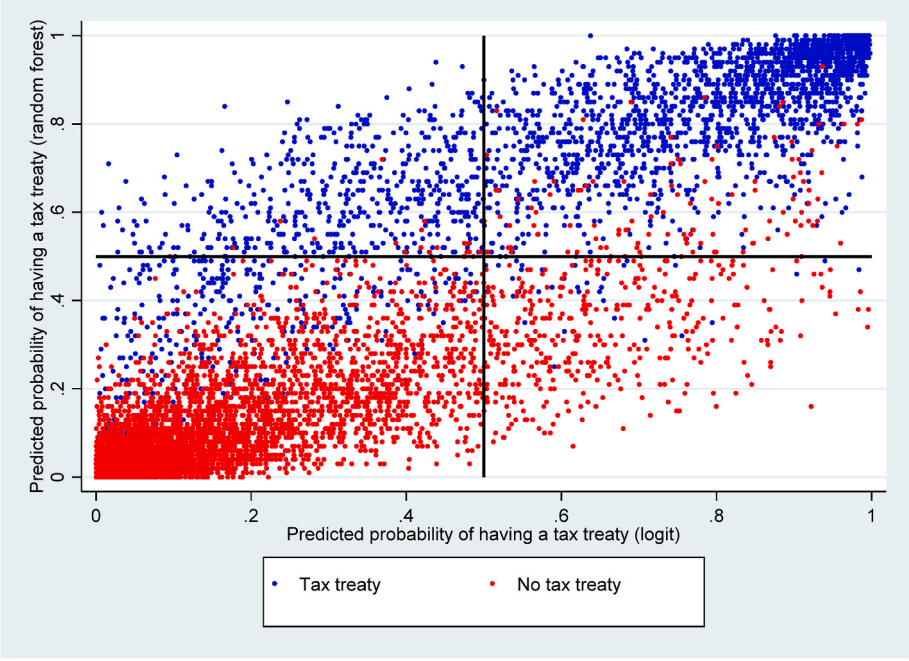


Fig. 15. Probability of having a tax treaty predicted by random forest vs. probability of having a tax treaty predicted by logit vs. actual tax treaty status.

Table 6
Hyperparameter selection outcome (10-fold cross-validation).

Method	Parameter 1	Parameter 2	Parameter 3	Cross-validation training accuracy	Cross-validation testing accuracy
Random forest	Optimal tree depth = 10	Optimal n. of splitting features = 5	Optimal n. of trees = 14	0.988	0.934
Classification tree	Optimal tree depth = 3			0.948	0.922
Boosting	Optimal learning rate = 0	Optimal n. of trees = 1	Optimal tree depth = 1	0.927	0.927
Nearest Neighbor	Optimal n. of nearest neighbors = 18	Optimal kernel function = distance		1	0.923
Regularized multinomial	Optimal penalization parameter = 0.01	Optimal elastic parameter = 0		0.927	0.927
Neural network	Optimal n. of neurons in layer 1 = 1	Optimal n. of neurons in layer 2 = 1	Optimal L2 penalization = 0	0.927	0.927
Support vector machine	Optimal C parameter = 0.01	Optimal GAMMA parameter = 0		0.945	0.927

classification command to estimate the accuracy of the models. The logit regression correctly classifies 84.97% of observations in the test sample, and the probit regression 84.99%. This implies that for the analysis of tax treaty formation random forest outperforms conventional econometric methods.

In order to compare the results of Table 4 with Fig. 10, we calculate fully standardized coefficients obtained from the Logit regression[16] and put them in relation to the highest coefficient to identify their importance for the model. In Fig. 14 we plot this against the results obtained in Fig. 10.[17] The further up, the more important is a variable in our machine learning algorithm, the further to the left, the more important it is in our logit regression. The line represents points where machine learning and logit exhibit identical importance, for every variable above (below) the line machine learning considers it more (less) important than logit.

We find that both machine learning and conventional econometrics identify the trade sum as the most important exogenous variables. A major importance of double tax treaties is the avoidance of double taxation for multinational corporations. Whereas machine learning quite sensibly identifies FDI – both the sum (point 4 in Fig. 14) and the difference (point 2) between these two countries, Logit (and Probit) point towards GDP (points 8 and 10) as a major explanation.

Fig. 15 plots random forest predictions against logit predictions against actual tax treaty status. Also graphically we see that random forest outperforms logit regression. Both of them are good in the quadrants I and III in predicting country pairs with tax treaties as having tax treaties (blue dots) and country pairs without tax treaties as not having tax treaties (red dots) respectively. However, as opposed to random forests, logit performs poor in the quadrants II and IV. It predicts many country pairs with tax treaties as not having tax treaties (the blue dots in the upper left quadrant) and many country pairs without tax treaties as having tax treaties (the red dots in the lower right quadrant), respectively. Indeed, the strength of machine learning lies in its predictability performance, demonstrated in this graph. We find only very few red dots in the upper half of the graph and very few blue dots in the lower half of the graph (which would be prediction errors of the random forest algorithm).

The preceding analysis should be interpreted in light of its constraints, primarily stemming from the limited variability in certain factors, notably distance. This constraint may exert an influence on the results by rendering the task more straightforward for the algorithms. In response to these potential limitations, we have undertaken a rigorous examination that specifically focuses on newly signed treaties. Although this refined approach may exacerbate data imbalances, it holds the promise of delivering more robust and enlightening outcomes, particularly with regard to predictive accuracy.

Given the low number of about 50 country pairs with new tax treaties per year, we apply the technique of random undersampling by randomly selecting 10% of country pairs without tax treaties in the train sample. After that, we have about 7.30% of country pairs with tax treaties and the rest without tax treaties in our train sample. Table 5 presents predictive accuracy of different algorithms with default parameters, whereas Table 6 conducts the hyperparameter selection.

Though support vector machine has a higher testing accuracy than other algorithms, it is based on a simple rule of classifying all country pairs as not having tax treaties. The random forest is the second best-performing algorithm. Its testing classification error rate is 0.009 in the default mode. When we do hyperparameter selection we have a CER of 0.012.[18]

When we look at which country pairs were likely to have new tax treaties in 2019 but did not have them yet based on the random forest prediction (predicted probability of having a tax treaty equal to or greater than 60% as in the main analysis above), these were Brazil-Germany, Brazil-United Kingdom, Brazil-United States, China-Dominican Republic, China-Myanmar, China-Samoa, Denmark-France, Germany-Qatar, and Saudi Arabia-United States. Most of these predictions already have a tax treaty relationship, which supports machine learning as a valid tool in tax treaty predictions. The Denmark-France tax treaty was signed in February 2022. Though there is no tax treaty between Brazil and the United States yet, there have been multiple attempts to negotiate it and the United States are considered the most important trade partner with whom Brazil does not have a tax treaty yet (Schoueri and Haddad, 2018). The Brazil-Germany tax treaty was terminated in 2005 (Dagnese, 2006) with a new treaty being under negotiation. A tax treaty between Brazil und the United Kingdom was signed in November 2022. Germany and Qatar have been negotiating a

[16] Logit performed slightly better than probit, but results are very similar.

[17] Note the axis are logarithmically scaled in order to avoid bunching around the origin. The axis represents absolute values, and only 9 variables are above 40 in one of the two dimensions (1–9 and 11).

[18] The value is obtained by inputting parameters optimized through cross-validation into the final test dataset. The prediction accuracy of a random forest algorithm without tuning may be greater than a random forest algorithm with tuning if the validation set is not representative of the general population. The model may become too specialized to the peculiarities of the validation set. The performance can also vary with different random seeds due to the randomness in selecting features and samples for building trees. It is possible that the untuned model got a "luckier" draw in terms of the subsets of the data it worked with, leading to better performance on the test set by chance. Another potential reason could lie in the data shift. If there is a significant shift in the distribution of the test data compared to the training and validation data, the model tuned on the latter might perform worse. An untuned model, being less specialized, might accidently be more robust to such shifts. Especially, it is to consider that given the low number of country pairs with a tax treaty in the whole dataset, their number in the cross-validation datasets may be even lower.

tax treaty.(See[fn19])

7. Conclusion

The paper analyzed country pairs that have tax treaties and country pairs that do not have tax treaties. For this, it applied novel machine learning techniques. Instead of relying on a theoretical model, it let the data speak, which is reasonable given the complex nature of the decision to enter into a tax treaty. A wide set of gravity variables was used to train the machine to distinguish between country pairs with tax treaties and country pairs without tax treaties. The year 2018 was chosen as the year to train the machine. In total, nine machine learning algorithms were trained and then tested using the 2019 data to estimate their predictive power. The random forest algorithm was selected as the one with the lowest testing classification error rate and thus the highest predictive power. The random forest was also found to outperform the conventional logit and probit regressions.

59 country pairs were identified that should have had tax treaties in year 2019 based on their features but had not had a tax treaty. 31 have already started or completed the negotiation process, whereas only 19 have to our knowledge not yet initiated a negotiation. Countries/regions with the highest number of predicted new tax treaties are Germany (9), Saudi Arabia (8), Brazil (7), Myanmar (7), and Hong Kong (6). All identified country pairs were then investigated in terms of their current tax treaty status.

Among the countries that have more than one missed opportunity for negotiation are Algeria, Brazil, China, Cyprus, Germany, Greece, the Netherlands, Myanmar, Saudi Arabia and Ghana. For policymakers in these countries in general, and their respective negotiation teams, these potential treaties present a clear opportunity to improve their treaty policy. For neighboring counties (such as France in the case of Germany and Belgium in the case of the Netherlands), these potential treaties pose a threat to their treaty network and tax policy. They may want to check whether they already have a treaty with respect to potential partners of their neighbors. Given a predicted treaty between Germany and Jordan and Germany and Peru, France may want to check whether it should start negotiating with Peru, or whether it should start improving conditions in its existing treaty with Jordan, which was last amended in 2019.

This paper has given a clear guideline how machine learning algorithms can give policymakers a clear indication of a course of action. In particular, we have used a particular machine learning algorithm, namely random forests, to predict potential future tax treaties between country pairs, and have argued that this gives a clear indication for national treaty negotiators on which treaty policy to follow.

Whereas the primary emphasis of this paper lies in the examination of country pairs with existing tax treaties as opposed to those without, utilizing their current features as the basis for comparison, a compelling avenue for future research could involve delving into the prediction of new tax treaties. The robustness check, which specifically considers newly signed tax treaties in the years 2018 and 2019, serves as a suggestive guidepost for this broader analytical framework.

Declaration of Competing interest

None.

Data availability

The authors do not have permission to share data.
Tax treaty formation (Original data) (Mendeley Data)

Appendix

Table A1
Variable description

Variable	Description	Data source
DTT	Dummy variable if countries have a signed tax treaty	Tax Treaties Explorer, IBFD Tax Research Platform
FDI sum	Sum of FDI stocks (in thousands current US$)	Calculated by the authors using IMF Coordinated Direct Investment Survey
FDI difference	Absolute difference of FDI stocks (in thousands current US$)	Calculated by the authors using IMF Coordinated Direct Investment Survey
Trade sum	Sum of trade flows (in thousands current US$)	Calculated by the authors using the CEPII Gravity Database. Original data source: IMF Direction of Trade Statistics
Trade difference	Absolute difference of trade flows (in thousands current US$)	Calculated by the authors using the CEPII Gravity Database. Original data source: IMF Direction of Trade Statistics
GDP sum	Sum of GDPs (current thousands US$)	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank's Development Indicators
GDP difference	Absolute difference of GDPs (current thousands US$)	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank's Development Indicators
GDP per capita sum	Sum of GDPs per capita (current thousands US$)	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank's Development Indicators
GDP per capita difference	Absolute difference of GDPs per capita (current thousands US$)	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank's Development Indicators
Population sum	Sum of populations (in thousands)	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank's Development Indicators
Population difference	Absolute difference of populations (in thousands)	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank's Development Indicators
Distance	Simple distance between most populated cities (km)	CEPII Gravity Database. Original data source: Geosphere R package
Contiguity	Dummy variable if countries are contiguous	CEPII Gravity Database. Original data source: ARCGIS's World Countries (Generalized) database

(*continued on next page*)

[19] We are only talking about publicly known status. At the same time, we cannot exclude that there are also non-public negotiations in a number of cases.

Table A1 (*continued*)

Variable	Description	Data source
Regional trade agreement	Dummy variable if the pair currently has an RTA	CEPII Gravity Database. Original data source: WTO's Regional Trade Agreements database
WTO membership	Variable is equal to 2 if both countries are WTO members, to 1 if one of the countries is WTO member, to 0 if none of the countries are WTO members	Calculated by the authors using the CEPII Gravity Database. Original data source: List of WTO members on WTO website
GATT membership	Variable is equal to 2 if both countries are GATT members, to 1 if one of the countries is GATT member, to 0 if none of the countries are GATT members	Calculated by the authors using the CEPII Gravity Database. Original data source: List of GATT members on WTO website
EU membership	Variable is equal to 2 if both countries are EU members, to 1 if one of the countries is EU member, to 0 if none of the countries are EU members	Calculated by the authors using the CEPII Gravity Database. Original data source: List of EU members on EU website
Entry cost difference	Absolute difference in cost of business start-up procedures (% of GNI per capita)	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank Development Indicators API
Entry time difference	Absolute difference in days required to start a business	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank Development Indicators API
Entry procedure difference	Absolute difference in number of start-up procedures to register a business	Calculated by the authors using the CEPII Gravity Database. Original data source: World Bank Development Indicators API
Common official or primary language	Dummy variable if countries share common official or primary language	CEPII Gravity Database. Original data source: CEPII's GeoDist
Common language spoken by at least 9% of the population	Dummy variable if countries share a common language spoken by at least 9% of the population	CEPII Gravity Database. Original data source: CEPII's GeoDist
Common religion	Religious proximity index	CEPII Gravity Database. Original data source: LaPorta et al. (1999):
Common colonizer post 1945	Dummy variable if countries share a common colonizer post 1945	CEPII Gravity Database. Original data source: CEPII's GeoDist
Colonial relationship post 1945	Dummy variable if countries are or were in colonial relationship post 1945	CEPII Gravity Database. Original data source: CEPII's GeoDist
Common legal origins before 1991	Dummy variable if countries share common legal origins before 1991	CEPII Gravity Database. Original data source: La Porta et al. (1999) and La Porta et al. (2008)
Common legal origins after 1991	Dummy variable if countries share common legal origins after 1991	CEPII Gravity Database. Original data source: La Porta et al. (1999) and La Porta et al. (2008)
Common legal origin change in 1991	Dummy variable if common legal origins changed in 1991	CEPII Gravity Database. Original data source: La Porta et al. (1999) and La Porta et al. (2008)
Colonial or dependency relationship ever	Dummy variable if pair ever was in colonial or dependency relationship (including before 1948)	CEPII Gravity Database. Original data source: Head et al. (2010), CIA World Factbook, Correlates of War Project (COW)
Same colonizer ever	Dummy variable if pair ever had the same colonizer (including before 1948)	CEPII Gravity Database. Original data source: Head et al. (2010), CIA World Factbook, Correlates of War Project

Table A2
Summary statistics (2018 training sample)

Variable	Mean	Standard deviation	Minimum	Maximum
DTT	1.313	0.464	1	2
FDI sum	1498688	17500000	0	898000000
FDI difference	1475117	13200000	0	467000000
Trade sum	2008019	18500000	0.002	1020000000
Trade difference	170652	1230614	0	58800000
GDP sum	1490000000	3400000000	1310732	34500000000
GDP difference	1290000000	3280000000	7955.162	20600000000
GDP per capita sum	37.403	30.571	0.800	199.472
GDP per capita difference	23.620	23.121	0.002	116.273
Population sum	117459.800	262939.600	149.200	2745347
Population difference	94712.490	251872.300	9.064	1392678
Distance	7232.871	4355.616	59.617	19812.040
Contiguity	0.024	0.152	0	1
Regional trade agreement	0.286	0.452	0	1
WTO membership	1.844	0.376	0	2
GATT membership	1.506	0.611	0	2
EU membership	0.435	0.564	0	2
Entry cost difference	23.673	33.607	0	200.300
Entry time difference	16.643	22.003	0	173.500
Entry procedure difference	3.167	2.402	0	14
Common official or primary language	0.151	0.358	0	1
Common language spoken by at least 9% of the population	0.151	0.358	0	1
Common religion	0.166	0.243	0	0.993
Common colonizer post 1945	0.087	0.282	0	1
Colonial relationship post 1945	0.011	0.106	0	1
Common legal origins before 1991	0.323	0.468	0	1
Common legal origins after 1991	0.376	0.484	0	1
Common legal origin change in 1991	0.097	0.296	0	1
Colonial or dependency relationship ever	0.016	0.125	0	1
Same colonizer ever	0.171	0.376	0	1

Table A3
Mean of the variables for country pairs with and without tax treaties in 2018

Variable/Mean value	Country pairs with tax treaties	Country pairs without tax treaties	Ha:diff <0	Ha:diff = 0	Ha:diff >0
FDI sum	4 575 965	95 834	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
FDI difference	4 501 843	95 308	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Trade sum	5 929 274	220 417	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Trade difference	459 436	39 003	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
GDP sum	2 270 000 000	1 130 000 000	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
GDP difference	1 830 000 000	1 050 000 000	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
GDP per capita sum	51	31	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
GDP per capita difference	28	22	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Population sum	170 562	93 252	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Population difference	137 832	75 055	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Distance	5222	8150	Pr(T < t) = 1.0000	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000
Contiguity	0.050	0.012	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Regional trade agreement	0.478	0.199	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
WTO membership	1.874	1.831	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
GATT membership	1.485	1.515	Pr(T < t) = 0.9851	Pr(T < t) = 0.0298	Pr(T < t) = 0.0149
EU membership	0.676	0.326	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Entry cost difference	10.722	29.577	Pr(T < t) = 1.0000	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000
Entry time difference	12.602	18.485	Pr(T < t) = 1.0000	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000
Entry procedure difference	2.939	3.270	Pr(T < t) = 1.0000	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000
Common official or primary language	0.1506	0.1509	Pr(T < t) = 0.5223	Pr(T < t) = 0.9554	Pr(T < t) = 0.4777
Common language spoken by at least 9% of the population	0.161	0.146	Pr(T < t) = 0.0325	Pr(T < t) = 0.0650	Pr(T < t) = 0.9675
Common religion	0.169	0.164	Pr(T < t) = 0.1882	Pr(T < t) = 0.3763	Pr(T < t) = 0.8118
Common colonizer post 1945	0.09	0.085	Pr(T < t) = 0.1771	Pr(T < t) = 0.3541	Pr(T < t) = 0.8229
Colonial relationship post 1945	0.028	0.004	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Common legal origins before 1991	0.328	0.321	Pr(T < t) = 0.2630	Pr(T < t) = 0.5259	Pr(T < t) = 0.7370
Common legal origins after 1991	0.369	0.379	Pr(T < t) = 0.8208	Pr(T < t) = 0.3583	Pr(T < t) = 0.1792
Common legal origins change in 1991	0.135	0.080	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Colonial or dependency relationship ever	0.037	0.006	Pr(T < t) = 0.0000	Pr(T < t) = 0.0000	Pr(T < t) = 1.0000
Same colonizer ever	0.175	0.169	Pr(T < t) = 0.2216	Pr(T < t) = 0.4433	Pr(T < t) = 0.7784

Table A4
List of countries/regions

Albania, Algeria, Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Comoros, Congo, Costa Rica, Cote d'Ivoire, Croatia, Cyprus, Democratic Republic of Congo, Denmark, Djibouti, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Estonia, Ethiopia, Fiji, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Grenada, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kiribati, Kuwait, Kyrgyz Republic, Laos, Latvia, Lebanon, Lesotho, Liberia, Libya, Luxembourg, Macedonia, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Mauritania, Mauritius, Mexico, Micronesia, Moldova, Mongolia, Morocco, Mozambique, Myanmar, Namibia, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, Norway, Oman, Pakistan, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Rwanda, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, Sao Tome and Principe, Saudi Arabia, Senegal, Seychelles, Sierra Leone, Singapore, Slovak Republic, Slovenia, Solomon Islands, South Africa, South Korea, Spain, Sri Lanka, Sudan, Suriname, Sweden, Switzerland, Tajikistan, Tanzania, Thailand, Togo, Tonga, Trinidad and Tobago, Tunisia, Turkey, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Vanuatu, Vietnam, Yemen, Zambia, Zimbabwe.

Table A5
Alternative performance metrics for classification of the test data

Method	Sensitivity	Precision	Specificity	F1-score	Area-under-the-curve
Random forest	0.869	0.951	0.979	0.908	0.983
Classification tree	0.829	0.834	0.921	0.831	0.875
Boosting	0.703	0.807	0.920	0.752	0.812
Nearest Neighbor	0.525	0.688	0.887	0.596	0.786
Regularized multinomial	0.579	0.742	0.905	0.650	0.743
Standard multinomial	0.579	0.742	0.905	0.650	0.743
Neural network	0.291	0.828	0.971	0.431	0.631
Support vector machine	0.313	0.778	0.958	0.447	0.636
Naive Bayes	0.228	0.708	0.956	0.345	0.596

$$\textbf{\textit{Sensitivity}} = \frac{\textit{true positives}}{\textit{true positives}+\textit{false negatives}}.$$

Sensitivity is applicable when we are intolerable towards false negatives. For example, in the case of diabetes diagnostics we would leave a diabetic person labelled healthy.

$$\textbf{\textit{Precision}} = \frac{\textit{true positives}}{\textit{true positives}+\textit{false positives}}.$$

Precision refers to the proportion of predicted positives that are actually positive. It measures how well a model can identify true positives. Precision is the metric of choice when the cost of false positives is significant. To exemplify, suppose we prefer receiving one extra spam email in our primary inbox than having a legitimate email flagged as spam.

$$\textbf{\textit{Specificity}} = \frac{\textit{true negatives}}{\textit{true negatives}+\textit{false positives}}.$$

Specificity refers to the proportion of actual negatives that are correctly identified as such by the machine learning model. It measures how well a model can identify true negatives. Specificity is the suitable parameter when the price of false positives is high. As an instance, let us consider a drug test, after which everyone who tests positive is sent to prison.

$$\textbf{\textit{F1}} - \textbf{\textit{score}} = \frac{2*\textit{sensitivity}*\textit{precision}}{\textit{sensitivity}+\textit{precision}}.$$

F1-score is the harmonic mean of precision and sensitivity and provides a balanced measure between the two metrics. It is useful when both precision and sensitivity are equally important.

The area under the curve (AUC) measures the overall performance of the classifier at all possible threshold values. The AUC ranges from 0 to 1, where a perfect classifier has an AUC of 1, and a completely random classifier has an AUC of 0.5. The AUC is calculated by plotting the receiver operating characteristic (ROC) curve. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) for different threshold values. The TPR is the proportion of true positive predictions among all actual positive cases, and the FPR is the proportion of false positive predictions among all actual negative cases. To calculate the AUC, the ROC curve is integrated using the trapezoidal rule. The area under the curve is then calculated by summing the areas of the trapezoids formed by adjacent points on the curve. The resulting value represents the overall performance of the classifier at all possible threshold values.

Table A6
Country pairs predicted to have tax treaties in 2019 and their current tax treaty status in 2023

Country/region A	Country/region B	Predicted probability of a tax treaty in 2019	Current status of a tax treaty in 2023 (IBFD, 2023; Orbitax, 2023)[18]
Signed			
Croatia	United States	0.62	Signed (December 2022)
Denmark	France	0.81	Signed (February 2022)
Brazil	United Kingdom	0.76	Signed (November 2022)
Brazil	Poland	0.67	Signed (September 2022)
Initialled			

(*continued on next page*)

Table A6 (*continued*)

Country/region A	Country/region B	Predicted probability of a tax treaty in 2019	Current status of a tax treaty in 2023 (IBFD, 2023; Orbitax, 2023)[18]
Croatia	Cyprus	0.77	Initialled
Greece	Japan	0.65	Initialled
Hong Kong	Turkey	0.85	Initialled
Under negotiation			
Albania	Slovak Republic	0.60	Under Negotiation
Argentina	South Korea	0.72	Under Negotiation
Australia	Bangladesh	0.66	Under Negotiation
Bahrain	India	0.63	Under Negotiation
Bangladesh	Hong Kong	0.69	Under Negotiation
Brazil	Germany	0.74	Under Negotiation
Brazil	Malaysia	0.81	Under Negotiation
Brazil	Malta	0.63	Under Negotiation
Brazil	Saudi Arabia	0.72	Under Negotiation
Brunei	Thailand	0.80	Under Negotiation
Chile	Germany	0.67	Under Negotiation
Cyprus	Israel	0.72	Under Negotiation
Denmark	Hong Kong	0.71	Under negotiation
Germany	Hong Kong	0.84	Under negotiation
Germany	Panama	0.64	Under negotiation
Germany	Saudi Arabia	0.93	Under negotiation
Germany	Senegal	0.65	Under Negotiation
Greece	Macedonia	0.75	Under negotiation
Hong Kong	Myanmar	0.62	Under negotiation
India	Nigeria	0.72	Under Negotiation
Italy	Peru	0.65	Under negotiation
Myanmar	Philippines	0.67	Under negotiation
Netherlands	Trinidad and Tobago	0.67	Under Negotiation
Nigeria	Switzerland	0.63	Under Negotiation
No tax treaty negotiations reported, but other kind of tax-related treaty signed or under negotiation			
Saudi Arabia	United States	0.65	Transport tax treaty signed (December 1999)
Italy	Nigeria	0.69	Transport tax treaty signed (February 1977)
Saudi Arabia	Thailand	0.75	Transport tax treaty signed (June 1994)
Indonesia	Saudi Arabia	0.71	Transport tax treaty signed (March 2001)
Canada	Panama	0.68	Transport treaty signed (February 2020)/Exchange of information agreement signed (March 2013)
Norway	Saudi Arabia	0.63	Air transport agreement under negotiation
Terminated			
Denmark	Spain	0.80	Terminated (January 2009)
Finland	Portugal	0.72	Terminated (January 2019)
Denmark	Moldova	0.66	Terminated (September 2003)
No tax treaty			
Albania	Cyprus	0.77	No tax treaty
Algeria	Cyprus	0.62	No tax treaty
Algeria	Greece	0.67	No tax treaty
Brazil	Hong Kong	0.73	No tax treaty
Canada	Ghana	0.61	No tax treaty
China	Ghana	0.64	No tax treaty
China	Myanmar	0.80	No tax treaty
Cote d'Ivoire	Netherlands	0.60	No tax treaty
Denmark	Saudi Arabia	0.67	No tax treaty
Finland	Saudi Arabia	0.67	No tax treaty
France	Myanmar	0.83	No tax treaty
Germany	Jordan	0.65	No tax treaty
Germany	Myanmar	0.60	No tax treaty
Germany	Peru	0.64	No tax treaty
Greece	Lebanon	0.60	No tax treaty
Guatemala	Italy	0.65	No tax treaty
Honduras	Mexico	0.64	No tax treaty
Japan	Myanmar	0.85	No tax treaty
Myanmar	Netherlands	0.86	No tax treaty

References

Abedin, M.Z., Hassan, M.K., Khan, I., Julio, I.F., 2022. Feature transformation for corporate tax default prediction: application of machine learning approaches. Asia Pac. J. Oper. Res. 39 (4), 2140017 https://doi.org/10.1142/S0217595921400170.

Abrell, J., Kosch, M., Rausch, S., 2022. How effective is carbon pricing?—a machine learning approach to policy evaluation. J. Environ. Econ. Manag. 112, 102589 https://doi.org/10.1016/j.jeem.2021.102589.

Alarie, B., Xue Griffin, B., 2022. Using machine learning to crack the tax code. Tax Notes Federal 661. https://ssrn.com/abstract=4033902.

Alarie, B., Niblett, A., Yoon, A.H., 2016. Using machine learning to predict outcomes in tax law. Can. Bus. LJ 58, 231. https://heinonline.org/HOL/P?h=hein.journals/canadbus58&i=249.

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., Salvestrini, V., 2018. Targeting with machine learning: an application to a tax rebate program in Italy. J. Econ. Behav. Organ. 156, 86–102. https://doi.org/10.1016/j.jebo.2018.09.010.

Athey, S., Imbens, G.W., 2019. Machine learning methods that economists should know about. Annual Review of Economics 11, 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Barthel, F., Neumayer, E., 2012. Competing for scarce foreign capital: spatial dependence in the diffusion of double taxation treaties. Int. Stud. Q. 56 (4), 645–660. https://doi.org/10.1111/j.1468-2478.2012.00757.x.
Basuchoudhary, A., Bang, J.T., Sen, T., 2017. Machine-learning Techniques in Economics: New Tools for Predicting Economic Growth. Springer. https://doi.org/10.1007/978-3-319-69014-8.
Battiston, P., Gamba, S., Santoro, A., 2024. Machine learning and the optimization of prediction-based policies. Technol. Forecast. Soc. Change 199, 123080. https://doi.org/10.1016/j.techfore.2023.123080.
Beyca, O.F., Ervural, B.C., Tatoglu, E., Ozuyar, P.G., Zaim, S., 2019. Using machine learning tools for forecasting natural gas consumption in the province of Istanbul. Energy Econ. 80, 937–949. https://doi.org/10.1016/j.eneco.2019.03.006.
Blonigen, B.A., Davies, R.B., 2004. The effects of bilateral tax treaties on US FDI activity. Int. Tax Publ. Finance 11 (5), 601–622. https://doi.org/10.1023/B:ITAX.0000036693.32618.00.
Braun, J., Zagler, M., 2018. The true art of the tax deal: evidence on aid flows and bilateral double tax agreements. World Econ. 41 (6), 1478–1507. https://doi.org/10.1111/twec.12628.
Bülhmann, P., van der Geer, S., 2011. Statistics for high dimensional data. Statistics (New York. Springer-Verlag, Berlin. https://doi.org/10.1007/978-3-642-20192-9.
Cao, X., 2010. Networks as channels of policy diffusion: explaining worldwide changes in capital taxation, 1998–2006. Int. Stud. Q. 54 (3), 823–854. https://doi.org/10.1111/j.1468-2478.2010.00611.x.
Cerulli, G., 2020. Machine learning using Stata/Python. https://arxiv.org/pdf/2103.03122.pdf.
Cerulli, G., 2021a. Improving econometric prediction by machine learning. Appl. Econ. Lett. 28 (16), 1419–1425. https://doi.org/10.1080/13504851.2020.1820939.
Cerulli, G., 2021b. C_ML_STATA: stata module to implement machine learning classification in Stata. https://ideas.repec.org/c/boc/bocode/s459055.html.
Cerulli, G., 2021c. R_ML_STATA: stata module to implement machine learning regression in Stata. https://ideas.repec.org/c/boc/bocode/s459054.html.
Cerulli, G., 2022. Machine learning using Stata/Python. STATA J. 22 (4), 772–810. https://doi.org/10.1177/1536867X221140944.
Chen, G., Wang, Z., 2021. Climbing to the top? How globalized competition for capital affects judicial independence. Stud. Comp. Int. Dev. 56 (4), 511–535. https://doi.org/10.1007/s12116-021-09345-6.
Chen, T.H., Chen, M.Y., Du, G.T., 2020. The determinants of bitcoin's price: utilization of GARCH and machine learning approaches. Comput. Econ. https://doi.org/10.1007/s10614-020-10057-7.
Chisik, R., Davies, R.B., 2004. Asymmetric FDI and tax-treaty bargaining: theory and evidence. J. Publ. Econ. 88 (6), 1119–1148. https://doi.org/10.1016/S0047-2727(03)00059-8.
Combes, P.P., Gobillon, L., Zylberberg, Y., 2022. Urban economics in a historical perspective: recovering data with machine learning. Reg. Sci. Urban Econ. 94, 103711 https://doi.org/10.1016/j.regsciurbeco.2021.103711.
Conte, M., Cotterlaz, P., Mayer, T., 2022. The CEPII gravity database. CEPII Working Paper N°2022-05. http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=8.
Daman, I.L.S., 2006. Why Saudi Arabia is expanding its treaty network. Int. Tax Rev. 1. https://www.proquest.com/docview/230206000.
Damgaard, J., Elkjaer, T., Johannesen, N., 2018. Piercing the veil. Finance Dev. 55 (2) https://doi.org/10.5089/9781484357415.022.
Dagnese, N., 2006. Is Brazil developed-termination of the Brazil-Germany tax treaty. Intertax 34, 195. https://doi.org/10.54648/taxi2006030.
Delogu, M., Lagravinese, R., Paolini, D., Resce, G., 2024. Predicting dropout from higher education: evidence from Italy. Econ. Modell. 130, 106583 https://doi.org/10.1016/j.econmod.2023.106583.
De Roux, D., Perez, B., Moreno, A., Villamil, M.D.P., Figueroa, C., 2018. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 215–222. https://doi.org/10.1145/3219819.3219878.
Droste, M., 2020. Pylearn. https://github.com/mdroste/stata-pylearn.
Elsayyad, M., 2012. Bargaining over tax information exchange. Working Paper of the Max Planck Institute for Tax Law and Public Finance No. 2012-02. https://doi.org/10.2139/ssrn.2012593.
Evers, M., 2013. Tracing the origins of The Netherlands' tax treaty network. Intertax 41 (6/7). https://doi.org/10.54648/taxi2013033.
Galindo, J., Tamayo, P., 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Comput. Econ. 15 (1), 107–143. https://doi.org/10.1023/A:1008699112516.
Ghoddusi, H., Creamer, G.G., Rafizadeh, N., 2019. Machine learning in energy economics and finance: a review. Energy Econ. 81, 709–727. https://doi.org/10.1016/j.eneco.2019.05.006.
Gogas, P., Papadimitriou, T., 2021. Machine learning in economics and finance. Comput. Econ. 57 (1), 1–4. https://doi.org/10.1007/s10614-021-10094-w.
Hearson, M., 2018. When do developing countries negotiate away their corporate tax base? J. Int. Dev. 30 (2), 233–255. https://doi.org/10.1002/jid.3351.
Hearson, M., 2021. Tax treaties explorer [Online database]. Brighton: International Centre for Tax and Development (ICTD). https://www.treaties.tax.
Hines Jr, J.R., 2010. Treasure islands. J. Econ. Perspect. 24 (4), 103–126. https://doi.org/10.1257/jep.24.4.103.
Hong, S., 2018. Tax treaties and foreign direct investment: a network approach. Int. Tax Publ. Finance 25 (5), 1277–1320. https://doi.org/10.1007/s10797-018-9489-0.
Hull, I., Grodecka-Messi, A., 2022. Measuring the impact of taxes and public services on property values: a double machine learning approach. arXiv preprint arXiv:2203.14751. https://doi.org/10.48550/arXiv.2203.14751.
IBFD, 2023. Tax research platform. https://research.ibfd.org/.
IMF, 2023. Coordinated direct investment survey. https://data.imf.org/?sk=40313609-F037-48C1-84B1-E1F1CE54D6D5.
Ippolito, A., Lozano, A.C.G., 2020. Tax crime prediction with machine learning: a case study in the municipality of São Paulo. ICEIS 1, 452–459. https://doi.org/10.5220/0009564704520459.
Joint Committee on Taxation, 2004. Explanation of proposed income tax treaty between the United States and Japan. https://www.congress.gov/108/cprt/JPRT91693/CPRT-108JPRT91693.pdf.
Kasy, M., 2018. Optimal taxation and insurance using machine learning—sufficient statistics and beyond. J. Publ. Econ. 167, 205–219. https://doi.org/10.1016/j.jpubeco.2018.09.002.
Kearney, 2022. Kearney foreign direct investment confidence index. https://www.kearney.com/foreign-direct-investment-confidence-index/2022-full-report.
Khan, M.Y., Qayoom, A., Nizami, M.S., Siddiqui, M.S., Wasi, S., Raazi, S.M.K.U.R., 2021. Automated prediction of Good Dictionary EXamples (GDEX): a comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques. Complexity. https://doi.org/10.1155/2021/2553199.
Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction policy problems. Am. Econ. Rev. 105 (5), 491–495. https://doi.org/10.1257/aer.p20151023.
KPMG, 2022. United Kingdom – double taxation convention with Brazil signed, not yet in force. https://home.kpmg/xx/en/home/insights/2022/12/flash-alert-2022-212.html.
Leung, C., Unteroberdoerster, O., 2008. Hong Kong SAR as a Financial Center for Asia: Trends and Implications. IMF. Working Paper, 08/57. https://ssrn.com/abstract=1112159.
Ligthart, J.E., Vlachaki, M., Voget, J., 2011. The Determinants of Double Tax Treaty Formation. Unpublished manuscript. https://www.tax.mpg.de/fileadmin/user_upload/LigthartVlachakiVoget2011cropped.pdf.
Lopez-Cariboni, S., Cao, X., 2015. Import competition and policy diffusion. Polit. Soc. 43 (4), 471–502. https://doi.org/10.1177/0032329215602888.
Lu, X.H., Mamiya, H., Vybihal, J., Ma, Y., Buckeridge, D.L., 2019. Application of machine learning and grocery transaction data to forecast effectiveness of beverage taxation. Medinfo 248–252. https://doi.org/10.3233/shti190221.
Masrom, S., Rahman, R.A., Mohamad, M., Rahman, A.S.A., Baharun, N., 2022. Machine learning of tax avoidance detection based on hybrid metaheuristics algorithms. IAES Int. J. Artif. Intell. 11 (3), 1153–1163. https://doi.org/10.11591/ijai.v11.i3.pp1153-1163.
Milner, C., Berg, B., 2017. Tax analytics artificial intelligence and machine learning–level 5. PwC Advanced Tax Analytics & Innovation. https://www.pwc.no/no/publikasjoner/Digitalisering/artificial-intelligence-and-machine-learning-final1.pdf.
Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. J. Econ. Perspect. 31 (2), 87–106. https://doi.org/10.1257/jep.31.2.87.
Orbitax, 2023. Orbitax International Tax Platform. https://www.orbitax.com/.
Paolini, D., Pistone, P., Pulina, G., Zagler, M., 2016. Tax treaties with developing countries and the allocation of taxing rights. Eur. J. Law Econ. 42 (3), 383–404. https://doi.org/10.1007/s10657-014-9465-9.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html.
Petkova, K., 2021. Withholding tax rates on dividends: symmetries versus asymmetries or single-versus multi-rated double tax treaties. Int. Tax Publ. Finance 28 (4), 890–940. https://doi.org/10.1007/s10797-020-09637-y.
Petkova, K., Stasio, A., Zagler, M., 2020. Bilateral tax competition and regional spillovers in tax treaty formation. WU International Taxation Research Paper Series. https://doi.org/10.2139/ssrn.3567791, 2020-04).
Poulakias, K., 2021. Artificial intelligence and job automation: an EU analysis using online job vacancy data. CEDEFOP Working Paper Series (6). https://euagenda.eu/upload/publications/6206-en.pdf.
Rixen, T., Schwarz, P., 2009. Bargaining over the avoidance of double taxation: evidence from German tax treaties. Finanzarchiv/Public Finance Analysis 442–471. http://www.jstor.org/stable/40913238.
Schapire, R.E., 1990. The strength of weak learnability. Mach. Learn. 5 (2), 197–227. https://doi.org/10.1007/BF00116037.
Schapire, R.E., 2003. The Boosting Approach to Machine Learning: an Overview. Nonlinear Estimation and Classification, pp. 149–171. https://doi.org/10.1007/978-0-387-21579-2_9.
Schoueri, L.E., Haddad, G.L., 2018. Time for US-Brazil tax treaty. Fla. Tax Rev. 22, 885. https://heinonline.org/HOL/LandingPage?handle=hein.journals/ftaxr22&div=29&id=&page=.
Scikit-learn, 2022. Supervised learning. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning.
Shor, R., Weldon, R., 2009. Ruby and sapphire production and distribution: a quarter century of change. Gems Gemol. 45 (4), 236–259. https://www.gia.edu/doc/Ruby-and-Sapphire-Production-and-Distribution.pdf.
Soybilgen, B., Yazgan, E., 2020. Nowcasting US GDP using tree-based ensemble models and dynamic factors. Comput. Econ. https://doi.org/10.1007/s10614-020-10083-5.
Tanty, R., Desmukh, T.S., 2015. Application of artificial neural network in hydrology—a review. Int. J. Eng. Technol. Res 4, 184–188. https://doi.org/10.17577/IJERTV4IS060247.
Thrall, C., 2021. Treaty diplomacy and the global firm. https://www.calvinthrall.com/assets/treaty_regimes_IPES.pdf.

Tian, Y., Zhang, Y., 2022. A comprehensive survey on regularization strategies in machine learning. Inf. Fusion 80, 146–166. https://doi.org/10.1016/j.inffus.2021.11.005.

US Geological Survey, 2022. Distribution of rare earths production worldwide as of 2021, by country. [Graph]. Statista. https://www.statista.com/statistics/270277/mining-of-rare-earths-by-country/.

Varian, H.R., 2014. Big data: new tricks for econometrics. J. Econ. Perspect. 28 (2), 3–28. https://doi.org/10.1257/jep.28.2.3.

Xiao, G., 2004. People's Republic of China's round-tripping FDI: scale, causes and implications. ADBI Discussion Paper, 7. http://hdl.handle.net/10419/53496.

Yoon, J., 2020. Forecasting of real GDP growth using machine learning models: gradient boosting and random forest approach. Comput. Econ. https://doi.org/10.1007/s10614-020-10054-w.

Zhang, Q., Ni, H., Xu, H., 2023. Nowcasting Chinese GDP in a data-rich environment: lessons from machine learning algorithms. Econ. Modell. 122, 106204 https://doi.org/10.1016/j.econmod.2023.106204.

Zhang, Z., 2016. Introduction to machine learning: K-nearest neighbors. Ann. Transl. Med. 4 (11).

Zhang, T., Lin, W., Vogelmann, A.M., Zhang, M., Xie, S., Qin, Y., Golaz, J.C., 2021. Improving convection trigger functions in deep convective parameterization schemes using machine learning. J. Adv. Model. Earth Syst. 13 (5) https://doi.org/10.1029/2020MS002365.

Zheng, Y., Zheng, H., Ye, X., 2016. Using machine learning in environmental tax reform assessment for sustainable development: a case study of Hubei Province, China. Sustainability 8 (11), 1124. https://doi.org/10.3390/su8111124.

Zhou, X., Li, Y., 2022. Forecasting the COVID-19 vaccine uptake rate: an infodemiological study in the US. Hum. Vaccines Immunother. 18 (1), 2017216 https://doi.org/10.1080/21645515.2021.2017216.

Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G., Pineda, C., 2021. Identifying tax evasion in Mexico with tools from network science and machine learning. Corruption Networks 89–113. https://doi.org/10.1007/978-3-030-81484-7_6.