Journal of Physics:
Complexity

Identifying key products to trigger new exports: an explainable machine learning approach

To cite this article: Massimiliano Fessina *et al* 2024 *J. Phys. Complex.* **5** 025003

IOP Publishing J. Phys. Complex. 5 (2024) 025003 https://doi.org/10.1088/2632-072X/ad3604

Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
20 January 2023

REVISED
19 February 2024

ACCEPTED FOR PUBLICATION
20 March 2024

PUBLISHED
18 April 2024

PAPER

Identifying key products to trigger new exports: an explainable machine learning approach

Massimiliano Fessina[1], **Giambattista Albora**[2,3,*], **Andrea Tacchella**[2] **and Andrea Zaccaria**[2,4]

[1] IMT School for Advanced Studies, Lucca, Italy
[2] Centro Ricerche Enrico Fermi, Rome, Italy
[3] Sapienza University of Rome, Rome, Italy
[4] Institute for Complex Systems, CNR, UOS Sapienza, Rome, Italy
[*] Author to whom any correspondence should be addressed.

E-mail: alboragiambattista@gmail.com

Keywords: machine learning, complex networks, economic complexity, relatedness, country exports, feature importance

Supplementary material for this article is available online

Abstract

Tree-based machine learning algorithms provide the most precise assessment of the feasibility for a country to export a target product given its export basket. However, the high number of parameters involved prevents a straightforward interpretation of the results and, in turn, the explainability of policy indications. In this paper, we propose a procedure to statistically validate the importance of the products used in the feasibility assessment. In this way, we are able to identify which products, called *explainers*, significantly increase the probability to export a target product in the near future. The explainers naturally identify a low dimensional representation, the Feature Importance Product Space, that enhances the interpretability of the recommendations and provides out-of-sample forecasts of the export baskets of countries. Interestingly, we detect a positive correlation between the complexity of a product and the complexity of its *explainers*.

1. Introduction

The mechanisms underlying economic development [1] are among the most studied branches of economics since the work of Adam Smith [2]. However, the identification of its determinants remains an open problem, despite the flourishing of different models and interpretations [3]; in particular, standard theories, based in aggregated measures of production inputs, have limited capacity to predict growth and to recommend specific industrial policies [4]. The line of research based on the works of [5–7] moves from the presence of the so called capabilities, the set of endowments countries have and that permit their industrialization and developments. Capabilities are, in practice, hard both to define and measure, since in principle they could span from human capital, to infrastructures, government and so on. The solution proposed in [8] is to infer them from the export baskets, i.e. the diversification structure provided by the set of products exported by the country under investigation. This idea opened up the possibility to apply techniques and methodologies borrowed from physics and network science, which go under the name of economic complexity [9–12]. In particular, the approach discussed by Tacchella *et al* [11]. aims at building a synthetic measure of the Fitness of a country, which is able to forecast the GDP growth with a precision higher than the state of the art methodologies [13]. However, this approach provides a *global* picture of the country, while a more detailed analysis is often needed in order to provide specific industrial recommendations [14, 15]. In this perspective, a number of papers built networks whose nodes are products and links are given by their similarity, proxied by their co-occurrences in the export baskets of countries [10, 16, 17]. In such a way, two products can be defined as close in the sense that they share many of the capabilities needed in order to export them in a competitive way. Co-occurrences based approaches have however a low predictive performance, and this fact favors machine learning approaches as better tools to measure relatedness both at country [18–20] and firm level [21, 22]. In [16, 23, 24], the authors proposed approaches to explicitly model the relationship among

products, capabilities, and development. These frameworks naturally lead to the concepts of product progression [16, 19, 25] and arrow of development [26]: the relationship between products is often not undirected, or symmetric, as in the product space [10], but *directed*: countries starts their development from simple products and gradually enter in more sophisticated markets, following well defined paths of development [16]. Obviously, the identification of the specific products enabling countries to competitively export a given target product is a key element to design industrial policies and strategic patterns of development. Despite the importance of this investigation, a specific analysis was missing because of the lack of suitable tools and algorithms able to successfully forecast the export of countries. However, thanks to the introduction of machine learning in the economic complexity analysis [18], the tools at disposal reached a maturity such that this investigation can start providing concrete and scientifically validated results. This is the aim of the present paper: to provide an algorithmic approach based on a highly predictive machine learning method to measure the importance of single products and sectors for a country to export a specific target product in a given amount of years. The link between starting and target products will be quantified by using the feature importance, a key tool of supervised machine learning algorithms that allows a clear interpretation of the outputs. Recently, the computer science community felt the necessity to provide tools (such as Shapley values) to increase the interpretability of machine learning models; this led to a number of theoretical results and practical investigations [27–29]. For an application of Shapley values to economic complexity-related issues, see [30, 31].

2. Results

2.1. The predictive framework

Our aim is to understand which products enable a country to export a given target product. To do so, we investigate the mechanisms underlying a machine learning based prediction approach [18]. Such approach considers the competitiveness level of each country's export on each product as *features* [19]: obviously, some products will be dominant in the forecast exercise, while others will be practically irrelevant. The *feature importance* [19, 32, 33] will be our statistically validated measure of the ability of a product to activate another product. It is obviously of key importance to adopt a framework which has an excellent forecasting power. The approach discussed here, based on the Random Forest (RF) algorithm [34], outperforms the networks of co-occurrences [18] as well as other supervised machine learning algorithms [19], also when other data typologies are considered [18, 21, 22]. Here we briefly summarize the predictive framework. Full details are provided in the methods section.

The predictive task is represented by the out-of-sample forecast of the appearance of new links in the country-product temporal network [19]. At a given year y, the network represents whether country c exports product p in a competitive way or not. Mathematically, it is identified by the adjacency matrix M whose elements are

$$M_{cp}(y) = \begin{cases} 1 & \text{if } RCA_{cp}(y) \geqslant 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where c is a country, p is a product, and y is the selected year. $RCA_{cp}(y)$ is the *Revealed Comparative Advantage* [35], and it quantifies the relative advantage in year y of country c in exporting product p (more details are provided in the section 5). Given the knowledge of the network in a certain time interval, the RF algorithm can be trained in an appropriate cross-validated framework to make out-of-sample predictions on the matrix $M_{cp}(y+\delta)$, starting from the knowledge of $M_{cp}(y)$, that is, which country exports which product. In particular, a different RF model is trained for each product, and the other products are used as inputs, or features; in such a way, the RF learns from the past which products are usually associated with the target product.

In the present study we cover the time span 1996–2018, and choose a time interval $\delta = 5$ years: the algorithm is trained on years 1996–2013 to make predictions on 2018. The number of countries is 169, while products are classified according to the Harmonized System (HS) 1992, which has a hierarchical structure: products can be aggregated in 97 different sectors (2-digit code level), or split into 5040 detailed products (6-digit code level). The size of the matrix M will change accordingly. In the section 5 we provide more details about the data and the construction of the predictive model.

2.2. Feature importances

The interpretation of the predictions provided by the RF starts with the quantification of the importance the algorithm assigns to each feature (i.e. product) during the training procedure. In our setting, the goal is to forecast whether a product p at the 6-digit level will be exported by a country in year $y+\delta$, knowing in which

of the 97 2-digit sectors the country is active in year y (always in the RCA sense). The 2-digit sectors are hence used as binary features, whose value is 1 if the RCA of the country on the 2-digit product is greater than 1, and the RCA is computed using the sum of the export volumes of the 6-digit products that belongs to the sector. The importance of a feature is a measure of how much the activity of a country in a sector (i.e. export or non-export) is informative in order to determine if it will export the 6-digit product p after δ years. The decision of using 2-digit sectors as features is due to the computational time needed for the construction of the model, which would have been prohibitive if we had used all 5040 products.

The quantification of feature importances is obtained using the *Gini importance* [36] (or *mean impurity decrease*), a Random Forest-specific measure assigning to the features importance values summing up to 1. The mathematical definition of Gini impurity is provided in the section 5. Starting from the raw values, we performed a suitable statistical validation procedure, computing the corresponding *p-values* and imposing a validation threshold of 95%: such validation is based on the computation of the *null importances*, i.e. the importance values the algorithm assigns to each variable after its association with the target vector is broken (see materials and methods for a detailed description of the procedure) [37]. Only the statistically validated importances are kept, while the others are put equal to zero. Hence we obtain, for each of the 5040 predicted products, a vector containing the validated importance measures for the 97 aggregate productive sectors. We call the products retaining a significant importance value *explainers*: these products enhance the probability of a country to competitively export the target product as they signal the presence of the capabilities needed for it. In figure 1 we report the barplots of the feature importances for the products 'Tobacco (not stemmed or stripped)' (code **240110**), 'Sports footwear' (code **640411**) and 'Vacuum cleaners' (code **850910**), showing the 10 most important and the 5 least important sectors. The colors represents whether the feature importance has been statistically validated (blue), or not (red). In all three cases we can notice how the explainers can be intuitively related to the products: e.g. the 2-digit sectors to which the 6-digit products belong are correctly recovered among the explainers (respectively, 'Tobacco and tobacco substitutes', code **24**, 'Footwear; gaiters and the like', code **64**, and 'Electrical machinery and equipment', code **85**). This represent a first qualitative test of the ability of the implemented methodology to recover significant correlations between productive sectors and products, as learned by Random Forest in its training procedure.

2.3. Feature importance product space

The Gini importance vectors can be interpreted as high-dimensional representations for the products, like word embeddings [38] in natural language processing [39]. Indeed, they contain information about the productive background that the Random Forest algorithm recognizes as necessary or highly predictive for their future export. Hence, the distance between such vectors can be used as a proxy for products' similarity: two products whose Gini importance vectors are close need a similar presence/absence pattern of capabilities in order to be competitively exported.

To test this hypothesis we projected the 97-dimensional vectors on a 2-dimensional continuous space, using t-SNE [40]: a popular dimensionality reduction technique used for visualizing high-dimensional data in a lower-dimensional space. In short, it models the pairwise similarities between high-dimensional data points and maps them to a lower-dimensional space, preserving local structure and revealing meaningful patterns or clusters. The result, which we call Feature Importance Product Space (FIPS), is reported in figure 2. Here, each dot represents a 6-digit product, and the colors correspond to ten aggregate macro-categories (see supplementary information section S3). The structure of the FIPS is heterogeneous, with clusters of products belonging to single categories, as for Agrifood and Textiles (left side of the plot) and regions with the superposition of different product categories, as in the right side of the plot, where there is a mixing of Machinery, Vehicles, Chemicals and Instruments. This differentiation can be traced back to the complexity of products making up different sectors: less sophisticated sectors tend to be more distinguishable, as they need few capabilities, and therefore share similarity patterns with a smaller set of other products (see Materials and methods and supplementary information section S1 for further analyses). On the contrary, high-complexity products share large portions of the respective production lines and supply chains [41]. To highlight the ability of the space to identify the similarity of products even if they originally belong to different productive categories, we pinpointed two small clusters: the first (upper-left side of the figure) groups products related to the fur manufacture; the second (lower-right side of the figure) puts together different typologies of products, all related to the spacecraft industry.

2.4. Predicting products' appearances with FIPS

The ability to make out-of-sample forecasting on the country-product network represents the natural field to test and compare the validity of relatedness measures [19]. Therefore, in order to quantify the goodness of the FIPS reconstruction and the amount of information it brings, we use it to predict the appearance of new products in $M_{cp}(2018)$, employing a density-based approach [10]. In other words we predict that countries

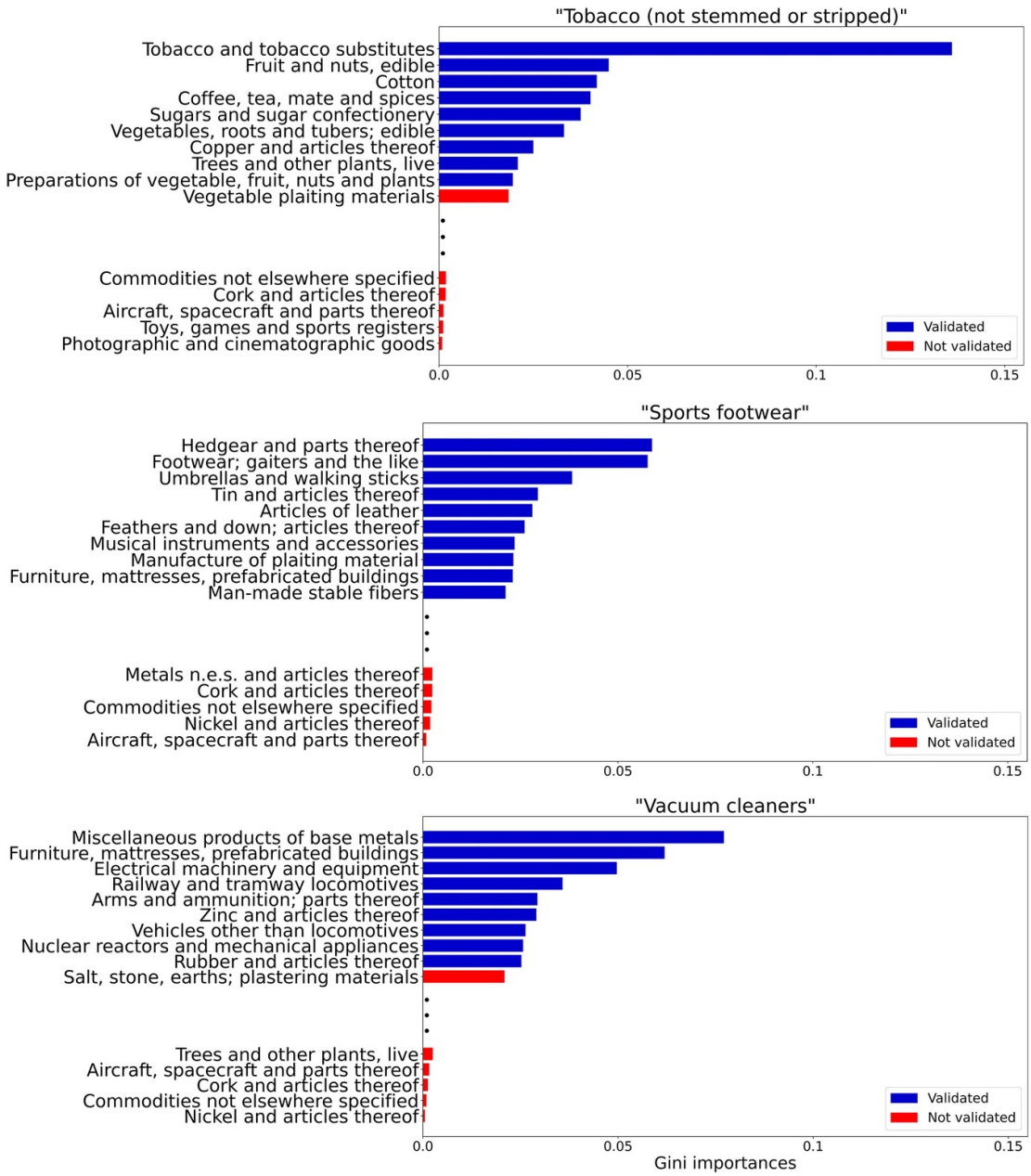

Figure 1. Feature importances barplots. Three examples of feature importances barplots; from top to bottom, 'Tobacco', 'Sports footwear', and 'Vacuum cleaners'. We report the top 10 and bottom 5 features: for all three products, the method validates a set of very reasonable features. Noticeably, in all three cases, the aggregated sectors to which the products belong are present among the explainers: 'Tobacco and tobacco substitutes' (code **24**) for 'Tobacco (not stemmed or stripped)' (code **240110**), 'Footwear; gaiters and the like' (code **64**) for 'Sports footwear' (code **640411**) and 'Electrical machinery and equipment (code **85**) for 'Vacuum cleaners' (code **850910**).

will become competitive in new products which are close in the FIPS space to other products in which the country is already competitive. Practically predictions on a single product, for every country, are based on the amount of already exported products, each weighted by its link with the target product. In table 1 we report the predictive performance of the FIPS, together with the performance of the Random Forest from which the FIPS was built, and with the temporal auto-correlation baseline represented by $RCA_{cp}(2013)$, both at the 6-digit and at the 2-digit aggregation level[5]. The RCA baseline involves utilizing the country's revealed comparative advantage value for the product in 2013 as an estimate of the probability that the country will export the product in 2018: for the 2-digit, the country's comparative advantage in sector s is attributed to all the 6-digit products p belonging to s. The rationale behind this approach is that it is more likely for a country

[5] In the remainder of the paper, if not otherwise specified, the generic expression RCA is referred to the 6-digit aggregation level.

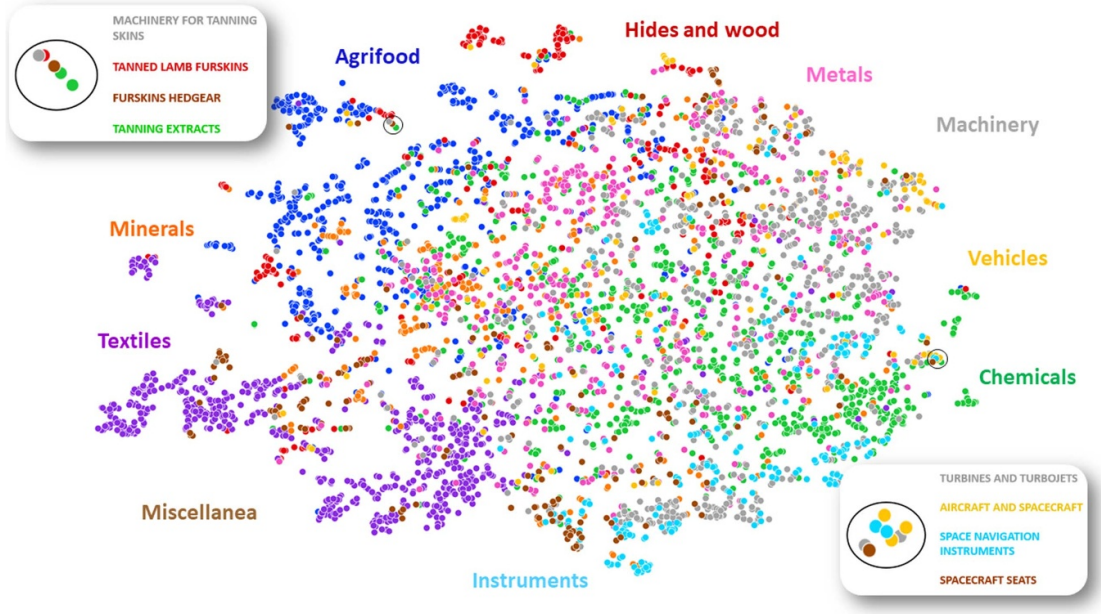

Figure 2. Feature Importance Product Space (FIPS). The FIPS is a low-dimensional representation of the feature importance vectors: each dot is a product, identified by its explainers. The colors correspond to ten productive macro-categories. The structure of the space is heterogeneous, with some relatively isolated categories of products (e.g. Textiles and Agrifood), and areas occupied by a mixing of different kinds of products (e.g. the right side of the figure, where there is a mixing of Machinery, Vehicles, Chemicals and Instruments). The two insets are zooms which testify the ability of the FIPS to group together similar products: the first (upper-left side) is composed of four different kind of products, all related to furskins; the second (lower-right side) groups together products belonging to the spacecraft industry.

Table 1. Comparison of prediction performances of FIPS, Random Forest and RCA baseline. The values of the performance metrics show that the FIPS performance is overall comparable with the Random Forest, showing higher values of Best F1 and mP@10, but lower values of AUC-ROC and AUC-PR: this result is very important, as it guarantees that the FIPS not only provides a fully interpretable predictive model in terms of products' similarity relationships, but retains the predictive power of the Random Forest it was built from. The RCA baseline has the highest Best F1 Score when built using the full 6-digit level data, while it provides the worst overall performance at the 2-digit level. The highest values of each indicator are in bold.

	Best F1 Score	AUC-ROC	mP@10	AUC-PR
Random Forest	0.033	**0.698**	0.041	**0.014**
FIPS	0.035	0.669	**0.047**	0.013
RCA - 6dig	**0.037**	0.592	0.039	0.012
RCA - 2dig	0.025	0.668	0.021	0.011

to export a product in the future if it already has a positive RCA for that product. The adopted performance metrics are (see Materials and methods for a detailed discussion):

- Best F1 Score: the F1-score [42], i.e. the harmonic mean of Precision and Recall [43], computed for the decision threshold that maximizes its value;
- AUC-ROC [44]: the area under the Receiving Operator Characteristic curve;
- mP@10: the average, over the countries, of the Precision score on the top 10 predicted products.
- AUC-PR [45]: the area under the precision-recall curve.

The performances have not been computed on the full matrix $M_{cp}(2018)$, but on the so called *activations*, i.e. those elements showing a value $RCA_{cp}(y) < 0.25$ for $y \in [1996–2013]$: as noted in previous works [18, 19], this guarantees that the measured performance is indicative of the actual ability of the models to forecast genuine economic development (i.e. the appearance of a new product in the export basket of a country), rather than relying on the strong temporal auto-correlation of the country-product network.

The scores show that the FIPS performs better than the original RF for both Best F1 Score and mP@10, while achieving a lower value of AUC-ROC and AUC-PR: this result is extremely relevant, as it implies that the FIPS has a forecasting power comparable to the Random Forest it was built from, while providing a clear interpretability of its predictions in terms of similarity relationships between products. Moreover, we stress that the AUC-ROC metrics is the least reliable, due to the strong class imbalance of the dataset (see [19] and

Table 2. Logistic regression carried out with FIPS and $RCA_{cp}(2013)$ to predict $M_{cp}(2018)$. The values of Pseudo R^2 show that the information carried by the FIPS and the RCA baseline are complementary, as the logistic regression trained on both models shows the highest value. This is confirmed by the performance metrics, as the latter shows a performance higher than both FIPS and RCA, when used individually (both directly and in a logistic regression setting) to make predictions; Random Forest anyway retains the highest AUC-ROC value. All performances are computed on the new products activations defined by $RCA_{cp}(y) < 0.25$ for $y \in [1996 - 2013]$. The asterisks indicate that all the coefficients are statistically validated within a 99.9% significance threshold.

	Logit model			Direct predictions		
	RCA	FIPS	FIPS + RCA	RCA	RF	FIPS
RCA	$9.015^{*} \pm 0.374$		$7.012^{*} \pm 0.406$			
FIPS		$4.782^{*} \pm 0.166$	$4.018^{*} \pm 0.177$			
Constant	$-5.223^{*} \pm 0.023$	$-5.315^{*} \pm 0.024$	$-5.391^{*} \pm 0.025$			
Pseudo R^2	0.014	0.019	**0.027**			
AUC-ROC	0.584	0.659	0.685	0.592	**0.698**	0.669
BestF1	0.035	0.034	**0.039**	0.037	0.033	0.035
mp@10	0.038	0.046	**0.053**	0.039	0.041	0.047
AUC-PR	0.011	0.013	**0.015**	0.012	0.014	0.013

Materials and methods). The RCA baseline at the 6-digit level, while trailing both FIPS and Random Forest in AUC-ROC, mP@10 and AC-PR, outperforms both models in Best F1 Score. This is due to the different granularity of the inputs, i.e. the 6-digit products for the former, and the 2-digit aggregated sectors for FIPS and RF, as confirmed by the superior performance of the latter models with respect to the 2-digit RCA baseline, which provides the worst overall performance (note that when the RF is trained at 6-digit it easily overcomes the 6-digit RCA baseline [18, 19]). However, the 6-digit RCA represents an important benchmark as it has been shown to perform substantially better than co-occurrence based approaches [18].

We expect, however, that the FIPS is uncovering fundamental capability-based explanations, that are sensibly different from the autocorrelation signal expressed by the RCA, and this cannot immediately be seen from the forecasting performance scores. In order to assess the additional information carried by the FIPS with respect to the temporal auto-correlation of the network, we decided to plug both the prediction score on $M_{cp}(2018)$ by FIPS and the 6-digit $RCA_{cp}(2013)$ as variables into a logistic regression whose dependent variable is the possible activation of a product. The logit model is trained on the *activations* ($RCA_{cp}(y) < 0.25$ for $y \in [1996–2013]$) in an appropriate cross-validated setting, to make out-of-sample predictions on $M_{cp}(2018)$ (see Materials and methods). The results, reported in table 2, confirm the validity of the information carried by the FIPS as complementary with respect to the network auto-correlation in two ways. First of all, the logit model trained on both FIPS and RCA has the highest value of Pseudo R^2; secondly, this model displays a better predictive performance with respect to both logit models trained on $RCA_{cp}(2013)$ and FIPS alone. We further compare it with the prediction accuracy provided by $RCA_{cp}(2013)$, RF, and FIPS alone (i.e. without being used as variables in a logistic regression), showing that the **FIPS + RCA** logit model has the highest Best F1 Score, and it trails only Random Forest for the AUC-ROC score.

2.5. Feature importance and products' complexity

Another key assessment of this study is the unveiling of a connection between the feature importance vector of a product and its complexity. The complexity of a product, defined applying the *Economic Fitness and Complexity* algorithm [11, 46] to the bipartite network country-product, is a non-monetary indicator related to the level of industrial sophistication needed to competitively export it on the global market. As such, we expect it to be connected to the nature of the *explainers* obtained for a product, as they represent the productive sectors recognized by our model as necessary for the future export of the product: the more complex a product, the more complex we expect the corresponding explainers to be. To measure the complexity of the features we applied the fitness and complexity algorithm to the bipartite network that connects countries with the 97 2-digit sectors. Since we train our Random Forest models using data in the time span 1996–2013, the complexities of products were computed as the average of the annual (log-) complexities in the same interval. The visualization of the average complexity of the validated features versus the complexity of the corresponding target products (figure 3) confirms this idea: more complex products need, on average, more complex features in order to be competitively exported. This finding confirms that the production lines of highly sophisticated products are deeply entangled among themselves [41, 47].

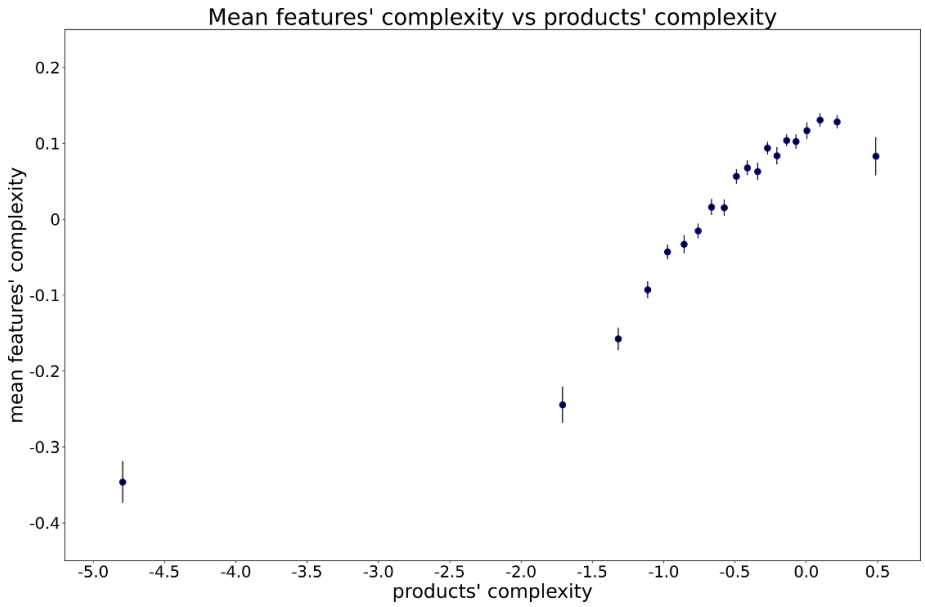

Figure 3. Mean complexity of explainers versus the mean complexity of target products. Non parametric regression of the mean complexity of features with respect to the complexity of the target products. The values of complexity are computed as averages of the corresponding logarithms in the time span 1996–2013. Target products are sorted by increasing complexity and then grouped into 20 bins of 252 products each, for which we compute average complexity and standard deviation.

3. Discussion

Relatedness [48] is a central topic of the economic complexity approach and a key element for investment decisions and policy makers [14, 15]. The idea is to empirically measure how close a country is to exporting a new product, that is to assess the feasibility of such a strategy. By comparing the predicting performances of different methodologies, recent studies [18, 19] showed that machine learning algorithms such as RF provide the state-of-the-art assessment of relatedness; here the features of this supervised machine learning approach are the products which are present or absent in the export basket of countries. The cost of a better prediction and relatedness assessment is, however, a reduced interpretability of the results, at least with respect to the traditional, network-based approaches [10, 16]. Nevertheless, having a visual representation of the diversification dynamics of countries, as well as knowing which products are the most relevant to activate (or to *explain*) the export of a new product is essential in order to inform industrial policies and to understand the different patterns of economic development. In this study, we address the problem of the black box nature of the RF algorithm by proposing a methodology to extract information on the relevance of each input feature (a 2-digit sector) as a predictor of the future export of each of the 5000 possible target products at 6 digits. The starting point is the construction of a predictive model for the possible future export of each target product, based on the training of a RF algorithm. We then apply a procedure to statistically validate the Gini Importance of the single input features; in this way we are able to identify the *explainers*, the key products needed by a country to competitively export a target product in the near future. The importance the algorithm assigns to each input feature for every target product can be arranged in a 97-dimensional feature importance vector, which represents a highly dimensional embedding of the about 5000 target products. By means of the *t-SNE* algorithm [49], we project such vectors on a 2-dimensional continuous space we call Feature Importance Product Space (FIPS). Here each point represents a product, and the closeness between points indicates that the corresponding products are similar in the sense that they share most of the explainers needed for their export. As such, this approach is closer to the theoretical approach discussed in the seminal papers by Teece *et al* [6, 50], in which the capability overlap between products is detected *a posteriori* by counting their co-occurrences, an approach known in the complexity field as the Product Space [10]. Here, instead, the proximity is assessed by comparing which *input* sectors are needed for the target products; similar explainers clearly imply similar capabilities.

4. Conclusions

The density-based approach employed in the Feature Importance Product Space (FIPS) has demonstrated its capability to forecast future exports, revealing that it provides better predictions than the Random Forest algorithm from which it is derived. Additionally, the integration of FIPS into a logistic regression model, alongside the Revealed Comparative Advantage (RCA) of countries on products, has not only affirmed the significance of FIPS as a predictor of future exports but also displayed superior performance over the strong benchmark model given by RCA itself. These results confirm the validity of our approach, highlighting that the FIPS not only retains the predictive power of the black-box algorithm it is based on but also enhances interpretability, a notable advantage over low performing network-based approaches. Importantly, the FIPS adeptly captures information about the complexity of products, successfully identifying the most sophisticated sectors within dense clusters and isolating less complex products. This understanding is crucial for characterizing the capabilities needed to be competitive in the export of complex products. In conclusion, our study acknowledges the limitation of using 97 2-digit aggregated sectors as features, which slightly reduces the predictive power of the RF compared to using the more granular 6-digit 5040 products. This limitation was a pragmatic choice due to computational constraints. However, future works will aim to optimize the model for more detailed feature analysis at the 4- and 6-digit levels, further enhancing the model's accuracy and applicability in the field of economic complexity and trade prediction.

5. Materials and methods

5.1. Data

The starting data used in this study is gathered by UN-COMTRADE and available upon subscription on the website https://comtrade.un.org. UN-COMTRADE provides the annual bilateral export flows between countries at the 6-digit product level. Products are classified according to the Harmonized Commodity Description and Coding System, in its 1992 version (HS-1992): each product is identified by a 6 digits code, where each couple of digits refers to a different aggregation level. The total number of products ranges from 97 at the 2 digit level (aggregated sectors), to 5040 at the 6 digit level (detailed products).

Since importers' and exporters' declarations not always coincide, a Bayesian reconciliation procedure [13] is performed on data, leading to the definition of the annual export matrices $E_{cp}(y)$. Each element corresponds to the export volume realized by country c, for product p, in year y. The total number of countries is 169, and the covered time span is 1996–2018.

Following the standard procedure in the economic complexity literature [10, 11], we compute the Revealed Comparative Advantage [35]:

$$RCA_{cp}(y) = \frac{E_{cp}/\sum_{p'} E_{cp'}}{\sum_{c'} E_{c'p}/\sum_{c'p'} E_{c'p'}}. \tag{2}$$

This economic indicator measures the ratio between the weight that the export of a product p has for country c and the weight it has on the global market. In this way, we can filter out the size effects of both countries and industrial sectors. Finally, imposing a threshold equal to $RCA_{cp} = 1$, distinguishing whether country c is a competitive exporter of product p in year y, we obtain the binary adjacency matrices $M_{cp}(y)$, as described in equation (1). The dimension of both the RCA and the M matrix is the number of countries on the rows (169) and the number of products on the columns, (5040 at the 6-digit level and 97 at the 2-digit level).

5.2. Random Forest

In order to forecast the export of countries, we train a supervised machine learning algorithm. In particular, we train one model for each target product; being the answer binary, we adopt a classification algorithm. RF [34] is an ensemble method based on the aggregation of several decision trees [51]: the final prediction of the algorithm is given by the average of the predictions made by the single trees.

The Random Forest has been shown [19] to be the top performing algorithm, together with XGBoost [52], for our predictive task, which is discussed in detail in the next section. We point out that XGBoost is practically unfeasible for the specific investigation discussed here because of the needed computational effort. Moreover, the extraction of the feature importances is much more direct in the case of Random Forest.

In this study we made use of the *Python* implementation provided by the library *scikit-learn*[6], which makes use of the *CART* version of the algorithm [36]. The hyperparameters [32] were set to their default values, a usual choice given the relative stability of the predictive performance [53–55].

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier

For a more detailed description of the use of the Random Forest for predicting countries' exports (training, overfitting, prediction power), we refer to [19, 21], in particular in the supplementary information of [19] it is shown that using the default values for the hyperparameters does not involve a significant worsening of the RF's performance.

5.2.1. Predictive model

The aim of the application of the Random Forest algorithm [34] to the country-product network is to build a predictive model able to forecast the export baskets of countries after δ years, given the knowledge of their present export baskets. This means predicting the structure of the network $M_{cp}(y+\delta)$ starting from $M_{cp}(y)$.

This is realized through the construction of a single model for each target product p', performing a binary classification task. Given the knowledge of the network in the time span $[y_0, y_f]$, such model is trained on the set:

- $X_{\text{train}} = \left\{M_{cp}(y), y \in [y_0, y_f - 2\delta]\right\}$
- $y_{\text{train}} = \left\{M_{cp'}(y), y \in [y_0 + \delta, y_f - \delta]\right\}$

and, in this process, learns which export baskets in X_{train} are associated to the countries exporting or not exporting p' (y_{train}). The test set is defined in a similar way:

- $X_{\text{test}} = M_{cp}(y_f - \delta)$
- $y_{\text{test}} = M_{cp'}(y_f)$.

In this way we make sure the test is performed on completely unforeseen data, and prevent the algorithm from having any information about the structure of the network in years $y > y_f - \delta$ during the learning phase. The data relative to different years is stacked together vertically: in this perspective each country in each year represents an observation, the export baskets for all products its features, and the possible export of p', δ years later, the corresponding class. Putting together the predictions provided for all products, we recover the full matrix of predictions whose elements $S_{cp}(y_f)$, can be tested against $M_{cp}(y_f)$. It is to be noted that the prediction on a single element $S_{cp}(y_f)$ is a probability value between 0 and 1, to be binarized with the choice of a threshold in order to be compared to the empirical element $S_{cp}(y_f)$.

So, for each product p', the model is trained to associate its possible future export from every country in year $y+\delta$, to the information about the respective export baskets of all products in year y. The rationale is that the algorithm will base its predictions upon learning the similarity patterns between different products, using different countries as different observations. To further explore the functioning of our predictive model based on Random Forest, we refer to [19] where the only difference is that, in the present study, we set the input data X at the 2 digit aggregation level: hence, for each of the 5040 6 digit products y, the input is represented by the export data about the 97 2 digit aggregated productive sectors.

5.2.2. Cross-validation

Given the strong temporal auto-correlation of the network [19], the knowledge of the present export basket of a country is very informative on its future export basket. So, in order to make sure that the predictions provided by the model are based solely on its learning of the correlations between products, rather than on its ability to recognize the country, we perform a 13-fold cross-validation procedure. The 169 countries are divided into 13 groups $\left\{C_k\right\}_{k=1}^{13}$ of 13 countries each. For each product, we then build 13 different models, where each one is trained on data about the 156 countries $c \notin C_k$ and is then used to make predictions for countries $c \in C_k$. In this way the predictions for every country are provided by a model that did not receive any information about the country itself. The supplementary information provides a schematic illustration to aid the reader in visualizing the delineation of the training and testing datasets, as well as the cross-validation process.

5.3. Feature importance

The directed link from a product p whose presence (or absence) enhances the likelihood that a general country exports also the target product p' is given by the feature importance, i.e. the relevance the RF algorithm attributes to each feature p in its predictive task. The construction of each decision tree in the forest is based on the recursive split of the observations that compose the training set, in terms of the corresponding values of the features [34]: starting from the root node (containing all the observations), each node considers a feature, and depending on the binary value of this feature, the observations are divided into two child nodes. The choice of the feature for each node is meant to maximize the decrease in *Gini impurity*,

a metrics measuring the impurity of a node as the compresence of observations belonging to both classes (i.e. 1 and 0), given by [51]:

$$G_j^m = \sum_{i=0,1} \hat{p}_{m,i}(1-\hat{p}_{m,i})$$

where m is the node, j the corresponding feature and $\hat{p}_{m,i}$ is the empirical frequency of observations in the node belonging to class i. The decrease in impurity realized by feature j on node m is then:

$$GD_j^m = G^m - f^1 G^{m,1} - f^2 G^{m,2}$$

where 1 and 2 indicate the two child nodes built in the split, and f the corresponding fractions of observations they receive. On a single tree t, being $N(j)$ the number of nodes to which feature j is attached, and V the total number of features, the decrease in impurity realized by feature j, i.e. its *Gini importance*, is equal to:

$$GI_j(t) = \frac{\sum_{m\in N(j)} GD_j^m(t)}{\sum_{j=1}^{V}\sum_{m\in N(j)} GD_j^m(t)}.$$

The *Gini importance* of a feature is then given by the average decrease in *Gini impurity* the feature realizes over the whole forest [36]:

$$GI_j = \frac{1}{T}\sum_{t=1}^{T} GI_j(t)$$

where T is the total number of trees.

5.3.1. Statistical validation procedure

Given the feature importance values, it is important to distinguish which features are actually informative for the algorithm, and which got a non-zero value because of spurious correlations in the dataset. We then implemented a statistical validation procedure similar to the one described in [37], in order to compute, for each feature importance, its corresponding p-value. The method is based on the reconstruction, for each feature importance, of the corresponding *null distribution*, i.e. a distribution of the importance values a feature is given by the algorithm under the hypothesis of independence between the feature itself and the response vector y_{train}.

The procedure works as follows:

1. For every product p', we train the Random Forest 50 times and compute the *Gini importance*, obtaining 50 vectors of feature importance $gi_n(p')$, $n = 1,\ldots,50$.
2. We then permute the response vector y_{train} 500 times, breaking its association with the feature, and recompute the *Gini importance* after every permutation. In this way we obtain 500 vectors of *null importance* $ni_m(p')$, $m = 1,\ldots,500$.
3. For each feature, we compare each of the 50 values of *Gini importance* with the 500 values of null importance: the corresponding *p-value* is computed as the fraction of 500 null importance values bigger than the *Gini importance* value. We then obtain, for each product, 50 vectors of *p-values* $pv_n(p')$, $n = 1,\ldots,50$.
4. We take the average vectors of *Gini importance*:

$$gi(p') = \frac{1}{50}\sum_{n=1}^{50} gi_n(p')$$

and we keep only the importance values of the features for which more than 95% of the *p-values* (i.e. at least 48 out of 50) are within the 95% significance threshold (i.e. $p < 0.05$), putting the others to 0.

In this way we obtain, for each product, a vector containing the 97 values of statistically validated feature importance, for the 97 features. The choice of the number of repetitions and permutations is consistent with the heavy computational cost involved. It has to be noted that point 1 and point 2 are carried out separately on each of the 13 folds of the *cross-validation* setting, and the corresponding values are averaged out. This computation required approximately 180 hours on a server with 20 cores

The method has been extensively tested on both low-dimensional [56] and high-dimensional [57] datasets, showing a great ability to filter out the non-informative features.

5.4. Feature importance product space

The feature importance vectors contain information about the productive sectors recognized by the Random Forest as important in order to competitively export the corresponding product 5 years later. Therefore, the distance between the vectors relative to two products can be seen as a natural proxy for their similarity, i.e. of the overlap of capabilities needed for their export. Then, in line with the *Continuous Projection Space* proposed in [18], we project these vectors on a 2-dimensional space, via the dimensionality reduction algorithm *t-SNE* [40]: on such space, which we call *Feature Importance Product Space* (FIPS), the distance between products is related to the distance between their original 97-dimensional vectors.

At this point, we can use the space to make out-of-sample forecasts on the activation of new exports after 5 years, by adopting the density-based approach explained in the following. This is a natural way to validate the FIPS idea and building procedure.

We first compute the matrix D of euclidean distances between products in the FIPS. Then we transform these distances into a similarity matrix B, where the similarity of two products p and p' is computed as:

$$B_{pp'} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{D_{pp'}}{\sigma}\right)^2\right] \tag{3}$$

where σ a free parameter.

Given this similarity matrix, following the economic complexity literature [10], we perform the prediction on $M_{cp}(y+\delta)$ by relating the likelihood of an activation to the scores defined by:

$$S_{cp}(y+\delta) = \frac{\sum_{p'} B_{pp'} M_{cp'}}{\sum_{p'} B_{pp'}} \tag{4}$$

i.e. for each country c the prediction on its future export of a product p is given by the sum on the the products it already exports, weighted by their similarity with p.

We build the FIPS on the Random Forest trained on data in years 1996–2013, and then used it to make out-of-sample forecasting on $M_{cp}(2018)$. We use of the *Python* implementation of *t-SNE* algorithm provided by the library *scikit-learn*[7].

5.4.1. Optimization of the parameters

The predictions provided by the FIPS (equation (4)) depend on two parameters: the *perplexity* value set for *t-SNE* and the standard deviation σ chosen for the gaussian weights (equation (3)). The former is a hyperparameter of the *t-SNE* algorithm, fixing the expected number of elements that will be grouped into each cluster [40]. The latter fixes the width of the gaussian distribution centered on each product to attribute the similarity weights to all the other products. The two parameters are then connected, as increasing the *perplexity* value will result in a denser FIPS, and hence even for small values of σ many products will get a high similarity score.

We therefore opt to combine them into a single parameter, which we call *average nearest neighbors*, computed empirically as the (average) number of neighbouring products contained within a circle of radius 3σ, centered on each product in the space. In practice, given a fixed value of *perplexity*, we look for the value of σ corresponding to integer values of *average nearest neighbors*, and then evaluate the performance of the FIPS, measured by *Best F1 Score* and *Mean Precision at 10*, as a function of this number. In figures 4 and 5 we show the trends of the two metrics for seven different values of *perplexity* ($P = 5, 10, 15, 25, 35, 50, 100$): in both cases curves corresponding to different *perplexity* values are quite close, with a peak for a value of average nearest neighbors around 70. We chose to set *perplexity* $= 10$, and the corresponding value $\sigma = 4.58$. The performance values reported in the section 2 are computed for these values of the parameters.

5.5. Logit model

To assess the additional information carried by the FIPS with respect to the temporal auto-correlation of the M matrices, we use the predictions provided by FIPS and $RCA(2013)$ as independent variables in a logistic regression for the probability of products' appearances in 2018, given by the equation:

$$S^{\text{logit}}_{cp}('18) = \alpha + \beta RCA_{cp}('13) + \gamma S^{\text{FIPS}}_{cp}('18) \tag{5}$$

[7] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

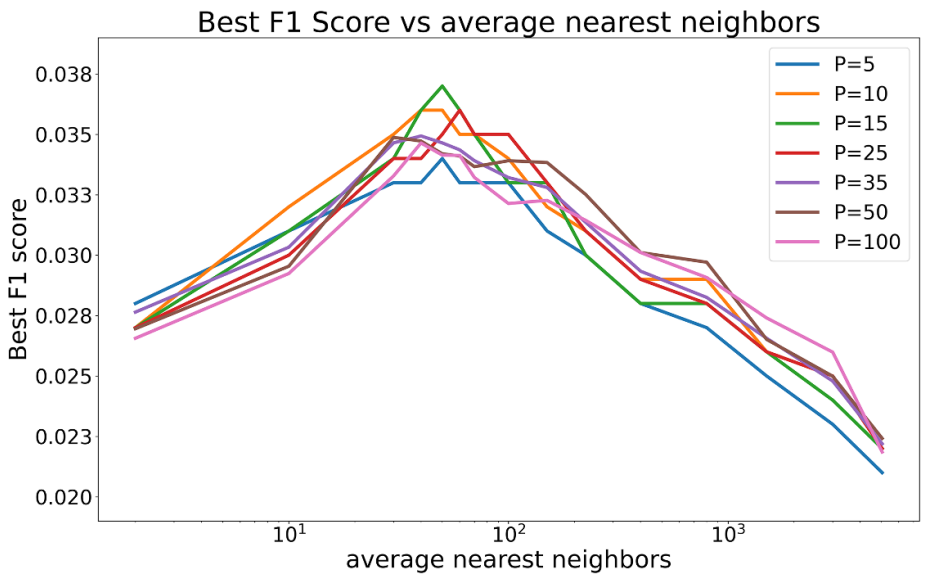

Figure 4. Best F1 Score versus average nearest neighbors for the FIPS. Curves corresponding to different values of *perplexity* are very close and show the same trend, with a clear peak for a number of average nearest neighbors equal to 70. Given the *perplexity* and the average nearest neighbors, the value of σ is fixed.

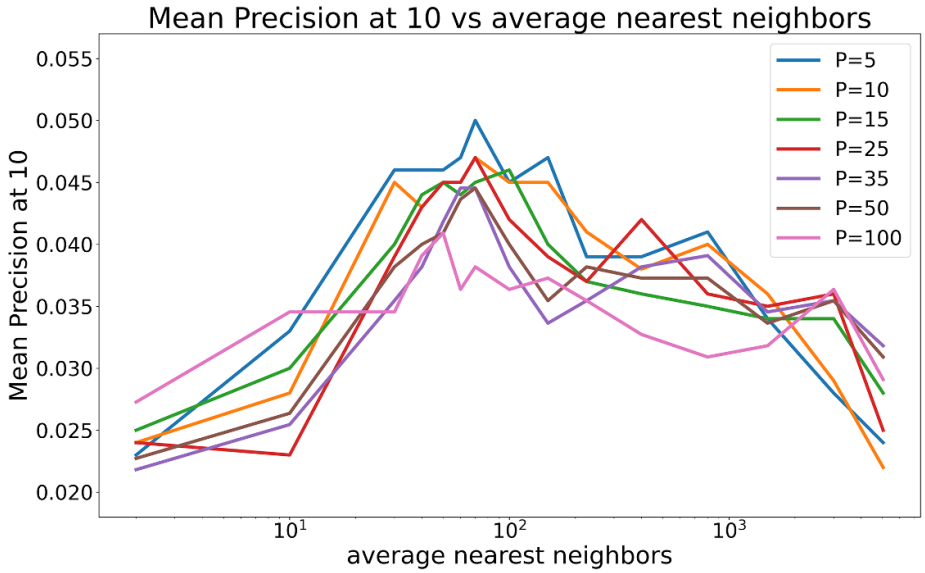

Figure 5. Mean Precision at 10 versus average nearest neighbors for the FIPS. Also in this case, curves corresponding to different values of *perplexity* are very close, and show a peak around 70 average nearest neighbors, but the trends are more noisy than for Best F1.

where $S_{cp}^{\text{FIPS}}('18)$ is the FIPS prediction for the element $m_{cp}('18)$. The prediction matrices $RCA(2013)$ and $S^{\text{FIPS}}(2018)$, as well as $M_{cp}(2018)$, are stacked into three vectors and the training set is built as:

- $X_{\text{train}} = (\vec{RCA}(2013), \vec{S}^{FIPS}(2018))$.
- $y_{\text{train}} = \vec{M}(2018)$.

The model is trained only on the *activations* (defined by $RCA_{cp}(y) < 0.25$ for $y \in [1996 - 2013]$, see [18]). In order to test the out-of-sample performance, we divide the training set into 13 subsets, following a cross-validation procedure: the predictions $\vec{S}_k(2018)$ for each group k ($k = 1, \ldots, 13$) are provided by a model trained on the remaining 12, and so can be tested against the corresponding elements $\vec{M}_k(2018)$.

The logistic regression was carried out using the *Logit* algorithm provided by the *Python* library *statsmodels*[8].

5.6. Performance metrics

To evaluate the predictive performances of the models, we made use of a series of evaluation metrics commonly used in Machine Learning. As already mentioned, the predictions $S_{cp}(2018)$ are probability values, to be binarized in order to compare them with the answers given by the matrix elements $M_{cp}(2018)$. In order to avoid the introduction of an arbitrary binarization threshold t, we opted for the use of '*threshold-free*' metrics, assessing the overall predictive performance of the models. Moreover, given the strong class imbalance of the dataset (the fraction of positive elements in the M_{cp} matrices oscillates around the 10% of the total elements in the covered time span, see [19]), we avoided metrics such as accuracy, awarding the correct individuation of true negatives (i.e. correct classification of elements $M_{cp}(2018) = 0$, which are often trivial). The chosen metrics are:

- **AUC-ROC**. The $AUC - ROC$, as suggested by the name (Area Under the Curve of the Receiving Operator Characteristic) [44, 58] measures the area under the Receiving Operator curve, i.e. the curve in the $TPR(t)$ *vs* $FPR(t)$ plane (respectively True Positive Rate and False Positive Rate, see [43]) obtained by varying the value of the binarization threshold t. Its value, ranging from 0 to 1, represents the probability that the classifier attributes an higher score to a positive element rather than to a negative one: $AUC - ROC = 1$ represents a perfect classifier, while $AUC - ROC = 0.5$ corresponds to a totally random classifier. It has been shown [45] that the $AUC - ROC$ is not fully reliable when the classifier is applied to an imbalanced dataset, which is our case (see [19]), as it tends to overestimate the actual accuracy of the predictions.
- **Mean Precision at *k***. The *precision* is defined as the ratio between the *true positives* (i.e. the positively classified elements that are actually positive) and all the positively classified elements [43]. We can define the *Precision at k* as the *precision* of the classifier on the k top-ranked elements, i.e. we classify the k elements with higher prediction scores as positives and then compute the corresponding *precision*. The *mean Precision at k* is obtained by computing the *Precision at k* for every country individually, and then taking the average over all countries. Since the most diversified countries tend to activate more products than the low and medium income ones, the averaging procedure allows to filter-out this effect, retaining an overall estimate of the classifier's performance. The value of k was set to 10.
- **Best F1 Score**. The *F1 Score* is defined as the harmonic mean of *precision* and *recall* [43]. Therefore it provides an estimate of the overall quality of the classifier, as it assumes an high value only if both *precision* and *recall* are high. Since these two quantities rely on the choice of a binarization threshold t, we adopted the *Best F1 Score*, i.e. the *F1 Score* computed for the value of t that maximizes it.
- **AUC-PR**. The *AUC-PR* measures the area under the curve drawn in the *precision(t)* vs *recall(t)* (see [43] for details) plane by varying the binarization threshold t. As such, it assesses the overall ability of the model to correctly classify positive elements. Differently from the *AUC-ROC*, the *AUC-PR* has been shown not to be affected by class imbalance [45].

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/mfessina/FIPS/tree/main. The starting database about the export volumes between countries is available from UN-COMTRADE (https://comtrade.un.org), upon subscription. The processed data used in this work is available from the authors upon reasonable request.

Acknowledgments

The authors would like to thank the CREF Project 'Complessità in Economia' and the European Union—Next Generation EU PRIN Project No. 20223W2JKJ 'WECARE'.

ORCID iD

Giambattista Albora https://orcid.org/0000-0001-8154-1607

[8] https://tedboy.github.io/statsmodels_doc/generated/generated/statsmodels.api.Logit.html.

References

[1] Acemoglu D 2012 Introduction to economic growth *J. Econ. Theory* **147** 545–50
[2] Smith A and Nicholson J S 1887 *An Inquiry Into the Nature and Causes of the Wealth of Nations* (T. Nelson and Sons)
[3] Helpman E 2009 *The Mystery of Economic Growth* (Harvard University Press)
[4] Barro R J 1989 Economic growth in a cross section of countries *Technical Report* (National Bureau of Economic Research)
[5] Penrose E and Penrose E T 2009 *The Theory of the Growth of the Firm* (Oxford University Press)
[6] Teece D J, Rumelt R, Dosi G and Winter S 1994 Understanding corporate coherence: theory and evidence *J. Econ. Behav. Organ.* **23** 1–30
[7] Sutton J 2012 *Competing in Capabilities: the Globalization Process* (Oxford University Press)
[8] Hausmann R, Hwang J and Rodrik D 2007 What you export matters *J. Econ. Growth* **12** 1–25
[9] Hidalgo C A and Hausmann R 2009 The building blocks of economic complexity *Proc. Natl Acad. Sci.* **106** 10570–5
[10] Hidalgo C A, Klinger B, Barabási A-L and Hausmann R 2007 The product space conditions the development of nations *Science* **317** 482–7
[11] Tacchella A, Cristelli M, Caldarelli G, Gabrielli A and Pietronero L 2012 A new metrics for countries' fitness and products' complexity *Sci. Rep.* **2** 723
[12] Sbardella A, Pugliese E, Zaccaria A and Scaramozzino P 2018 The role of complex analysis in modelling economic growth *Entropy* **20** 883
[13] Tacchella A, Mazzilli D and Pietronero L 2018 A dynamical systems approach to gross domestic product forecasting *Nat. Phys.* **14** 861–5
[14] Lin J, Cader M and Pietronero L 2020 *What African industrial development can learn from east Asian successes* 34852 (The World Bank Group)
[15] Pugliese E and Tacchella A 2020 Economic complexity for competitiveness and innovation: a novel bottom-up strategy linking global and regional capacities *Technical Report* (Joint Research Centre (Seville site))
[16] Zaccaria A, Cristelli M, Tacchella A and Pietronero L 2014 How the taxonomy of products drives the economic development of countries *PLoS One* **9** e113770
[17] Pugliese E, Cimini G, Patelli A, Zaccaria A, Pietronero L and Gabrielli A 2019 Unfolding the innovation system for the development of countries: coevolution of science, technology and production *Sci. Rep.* **9** 1–12
[18] Tacchella A, Zaccaria A, Miccheli M and Pietronero L 2021 Relatedness in the era of machine learning (arXiv:2103.06017)
[19] Albora G, Pietronero L, Tacchella A and Zaccaria A 2023 Product progression: a machine learning approach to forecasting industrial upgrading *Sci. Rep.* **13** 1481
[20] Che N X 2020 Intelligent export diversification: an export recommendation system with machine learning *Technical Report* (International Monetary Fund)
[21] Albora G and Zaccaria A 2022 Machine learning to assess relatedness: the advantage of using firm-level data *Complexity* **2022** 1–12
[22] Straccamore M, Pietronero L and Zaccaria A 2022 Which will be your firm's next technology? comparison between machine learning and network-based algorithms *J. Phys. Complex.* **3** 035002
[23] Tacchella A, Di Clemente R, Gabrielli A and Pietronero L 2016 The build-up of diversity in complex ecosystems (arXiv:1609.03617)
[24] Saracco F, Di Clemente R, Gabrielli A and Pietronero L 2015 From innovation to diversification: a simple competitive model *PLoS One* **10** e0140420
[25] Zaccaria A, Mishra S, Cader M Z and Pietronero L 2018 Integrating services in the economic fitness approach *World Bank Policy Research Working Paper*
[26] O'Clery N, Yıldırım M A and Hausmann R 2021 Productive ecosystems and the arrow of development *Nat. Commun.* **12** 1479
[27] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F and Pedreschi D 2018 A survey of methods for explaining black box models *ACM Comput. Surv.* **51** 1–42
[28] Holzinger A 2018 *2018 World Symp. on Digital Intelligence for Systems and Machines (DISA)* (IEEE) pp 55–66
[29] Belle V and Papantonis I 2021 Principles and practice of explainable machine learning *Front. Big Data* **4** 39
[30] Futagami K, Fukazawa Y, Kapoor N and Kito T 2021 Pairwise acquisition prediction with shap value interpretation *J. Financ. Data Sci.* **7** 22–44
[31] Gnecco G, Nutarelli F and Riccaboni M 2023 Matrix completion of world trade: an analysis of interpretability through shapley values *World Econ.* **46** 2707–31
[32] Géron A 2019 *Hands-on Machine Learning with Scikit-Learn, Keras and Tensorflow: Concepts, Tools and Techniques to Build Intelligent Systems* (O'Reilly Media, Inc.)
[33] Shalev-Shwartz S and Ben-David S 2014 *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press)
[34] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
[35] Balassa B 1965 Trade liberalisation and "revealed" comparative advantage[1] *Manch. Sch.* **33** 99–123
[36] Breiman L, Friedman J, Stone C and Olshen R 1984 *Classification and Regression Trees* (Chapman and Hall/CRC)
[37] Altmann A, Toloşi L, Sander O and Lengauer T 2010 Permutation importance: a corrected feature importance measure *Bioinformatics* **26** 1340–7
[38] Mikolov T, Chen K, Corrado G and Dean J 2013 Efficient estimation of word representations in vector space (arXiv:1301.3781)
[39] Jurafsky D 2000 *Speech & Language Processing* (Pearson Education India)
[40] Van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605
[41] Taglioni D and Winkler D 2016 *Making Global Value Chains Work for Development* (World Bank Publications)
[42] Van Rijsbergen C J 1974 Foundation of evaluation *J. Doc.* **30** 365–73
[43] Powers D M 2011 Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation *J. Mach. Learn. Technol.* **2** 1
[44] Hanley J A and McNeil B J 1982 The meaning and use of the area under a receiver operating characteristic (ROC) curve *Radiology* **143** 29–36
[45] Saito T and Rehmsmeier M 2015 The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets *PLoS One* **10** e0118432
[46] Tacchella A, Cristelli M, Caldarelli G, Gabrielli A and Pietronero L 2013 Economic complexity: conceptual grounding of a new metrics for global competitiveness *J. Econ. Dyn. Control* **37** 1683–91
[47] Angelini O and Di Matteo T 2018 Complexity of products: the effect of data regularisation *Entropy* **20** 814
[48] Hidalgo C A *et al* 2018 *Int. Conf. on Complex Systems* (Springer) pp 451–7

[49] Maaten L V D and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605
[50] Teece D J, Pisano G and Shuen A 1997 Dynamic capabilities and strategic management *Strateg. Manage. J.* **18** 509–33
[51] James G, Witten D, Hastie T and Tibshirani R 2013 *An Introduction to Statistical Learning* vol 112 (Springer)
[52] Chen T and Guestrin C 2016 *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 785–94
[53] Genuer R, Poggi J-M and Tuleau C 2008 Random forests: some methodological insights (arXiv:0811.3619)
[54] Fernández A, García S, Galar M, Prati R C, Krawczyk B and Herrera F 2018 *Learning From Imbalanced Data Sets* (Springer)
[55] Probst P, Wright M N and Boulesteix A-L 2019 Hyperparameters and tuning strategies for random forest *Wiley Interdiscip. Rev. Data Min. Knowl. Discovery* **9** e1301
[56] Hapfelmeier A and Ulm K 2013 A new variable selection approach using random forests *Comput. Stat. Data Anal.* **60** 50–69
[57] Janitza S, Celik E and Boulesteix A-L 2018 A computationally fast variable importance test for random forests for high-dimensional data *Adv. Data Anal. Classif.* **12** 885–915
[58] Bradley A P 1997 The use of the area under the ROC curve in the evaluation of machine learning algorithms *Pattern Recognit.* **30** 1145–59