Appendix A: Data

Table A1: Panel (A): List of predictors

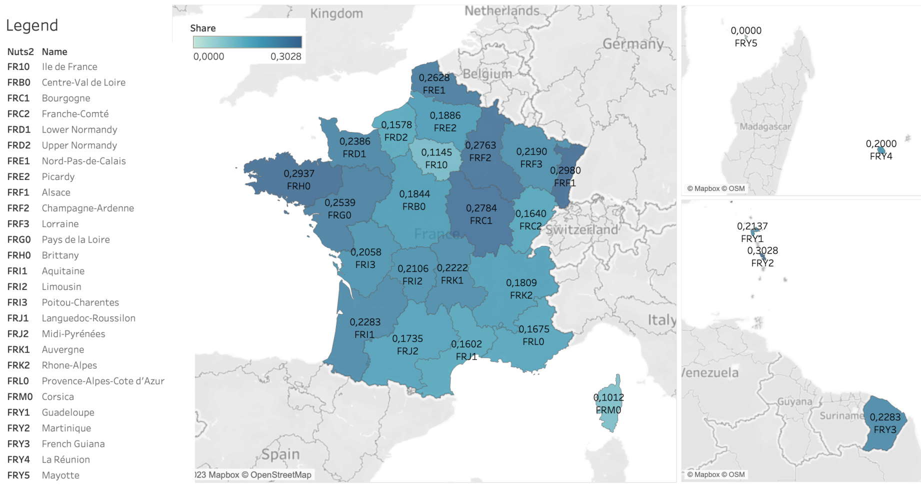
Variable	Description
Value Added, Depreciation, Creditors, Current Assets, Current liabilities, Non-current liabilities, Current ratio, Debtors, Operating Revenue Turnover, Material Costs, Costs of Employees, Taxation, Financial Revenues, Financial Expenses, Interest Paid, Number of Employees, Cash Flow, EBITDA, Total Assets, Fixed Assets, Intangible Fixed Assets, Tangible Fixed Assets, Shareholders' Funds, Long-Term Debt, Loans, Sales, Solvency Ratio, Working Capital	Original financial accounts expressed in euro.
Corporate Control	A binary variable equal to one if a firm belongs to a corporate group.
Dummy Patents	equal to 1 if the firm issued any patent, and 0 otherwise.
Consolidated Accounts	A binary variable equal to one if the firm consolidates accounts of subsidiaries
NACE rev. 2	A 2-digit industry affiliation following the European Classification
NUTS 2-digit	The region in which the company is located following the European classification.
Productive Capacity	It is an indicator of investment in productive capacity computed as $\textit{Fixed Assets}_t/(\textit{Fixed Assets}_{t-1}+\textit{Depreciation}_{t-1})$
Capital Intensity	It is a ratio between fixed assets and number of employees for the choice of factors of production.
Labour Productivity	It is a ratio between value added and number of employees for the average productivity of labor services.
Interest Coverage Ratio (ICR)	It is a ratio between EBIT and Interest Expenses, as yet another proxy of financial constraints as in Caballero et al. (2008).
TFP	It is the Total Factor Productivity of a firm computed as in Ackerberg et al. (2015).
Financial Constraints	It is a proxy of financial constraints as in Nickell and Nicolitsas (1999), calculated as a ratio between interest payments and cash flow

Table A1: Panel (B): List of predictors

Variable	Description
Markup	It an estimate of a firm's markup following De Loecker and Warzynski (2012).
ROA	It is a ratio of EBITDA on Total Assets for returns on assets.
Financial Sustainability	It is a ratio between Financial Expenses and Operating Revenues.
Size-Age	It is a synthetic indicator proposed by Hadlock and Pierce (2010), computed as (-0.737$\cdot$ $log(total\ assets)$)+(0.043 $\cdot$ $log(total\ assets))^2$ -(0.040 $\cdot$ age to catch the non-linear relationship between financial constraints, size and age.
Capital Adequacy Ratio	It is a ratio of Shareholders' Funds over Short and Long Term Debts.
Liquidity Ratio	A ratio between Current Assets minus Stocks and Current Liabilities.
Liquidity Returns	It is a ratio between Cash Flow and Total Assets
Regional Spillovers	It is a proxy proposed by Bernard and Jensen (2004) computed as a share of exporting plants out of total plants in a region.
Industrial spillovers	It is a proxy proposed by Bernard and Jensen (2004) computed as a share of exporting plants on total plants in a 2-digit industry.
External Economies of Scale	It is a proxy proposed by Bernard and Jensen (2004) computed as a share of exporting plants out of the total in an industry-region cell.
Size	Measure of firm size computed as (log of) number of employees.
Average Wage Bill	It is computed as (log of) costs of employees divided by number of employees.
Inward FDI	It is a binary variable with value 1 if the firm has foreign headquarters and 0 otherwise.
Outward FDI	It is a binary variable with value 1 if the firm has subsidiaries abroad and 0 otherwise.

Appendix B: Figures and Tables

Figure B1: Sample coverage: exporters by region



Note: Unitary shares indicate exporters on total firms in NUTS 2-digit regions.

Table B1: Sample coverage by industry

NACE rev. 2	code	Sample non-exporters	exporters	total	(%)	Population non-exporters	exporters	total	(%)
		(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Food products	10	13,057	1,429	14,486	0.254	49,153	2,135	51,288	0.293
Beverages	11	1,176	395	1,571	0.028	3,028	825	3,853	0.022
Textiles	13	919	389	1,308	0.023	4,278	798	5,076	0.029
Wearing apparel	14	1,060	336	1,396	0.024	8,813	881	9,694	0.055
Leather and related products	15	374	142	516	0.009	2,930	313	3,243	0.019
Wood and products of wood and cork	16	2,203	509	2,712	0.048	8,920	1,036	9,956	0.057
Paper and paper products	17	455	362	817	0.014	823	469	1,292	0.007
Printing and reproduction of recorded media	18	2,995	584	3,579	0.063	14,347	969	15,316	0.088
Coke and refined petroleum	19	17	14	31	0.001	-	-	25	0.0001
Chemicals and chemical products	20	958	705	1,663	0.029	1,388	1,127	2,515	0.014
Pharmaceutical products	21	151	148	299	0.005	93	159	252	0.001
Rubber and plastic products	22	1,436	931	2,367	0.042	1,780	1,425	3,205	0.018
Other non-metallic products	23	1,929	393	2,322	0.041	7,026	777	7,803	0.045
Basic metals	24	354	267	621	0.011	295	304	599	0.003
Fabricated metal prod., except machinery and equipment	25	8,135	2,540	10,675	0.187	14,557	3,903	18,460	0.106
Computer, electronic and optical products	26	965	605	1,570	0.028	1,304	991	2,295	0.013
Electrical equipment	27	789	495	1,284	0.023	1,321	727	2,048	0.012
Machinery and equipment	28	1,938	1,194	3,132	0.055	2,567	1,967	4,534	0.026
Motor vehicle, trailers and semi-trailers	29	748	424	1,172	0.021	1,119	516	1,635	0.009
Other transport equipment	30	330	186	516	0.009	847	260	1,107	0.006
Furniture	31	1,416	249	1,665	0.029	8,758	598	9,356	0.053
Other manufacturing	32	2,796	518	3,314	0.058	19,960	1,378	21,338	0.122
Total		44,201	12,815	57,016	1.00	153,307	21,558	174,890	1.00

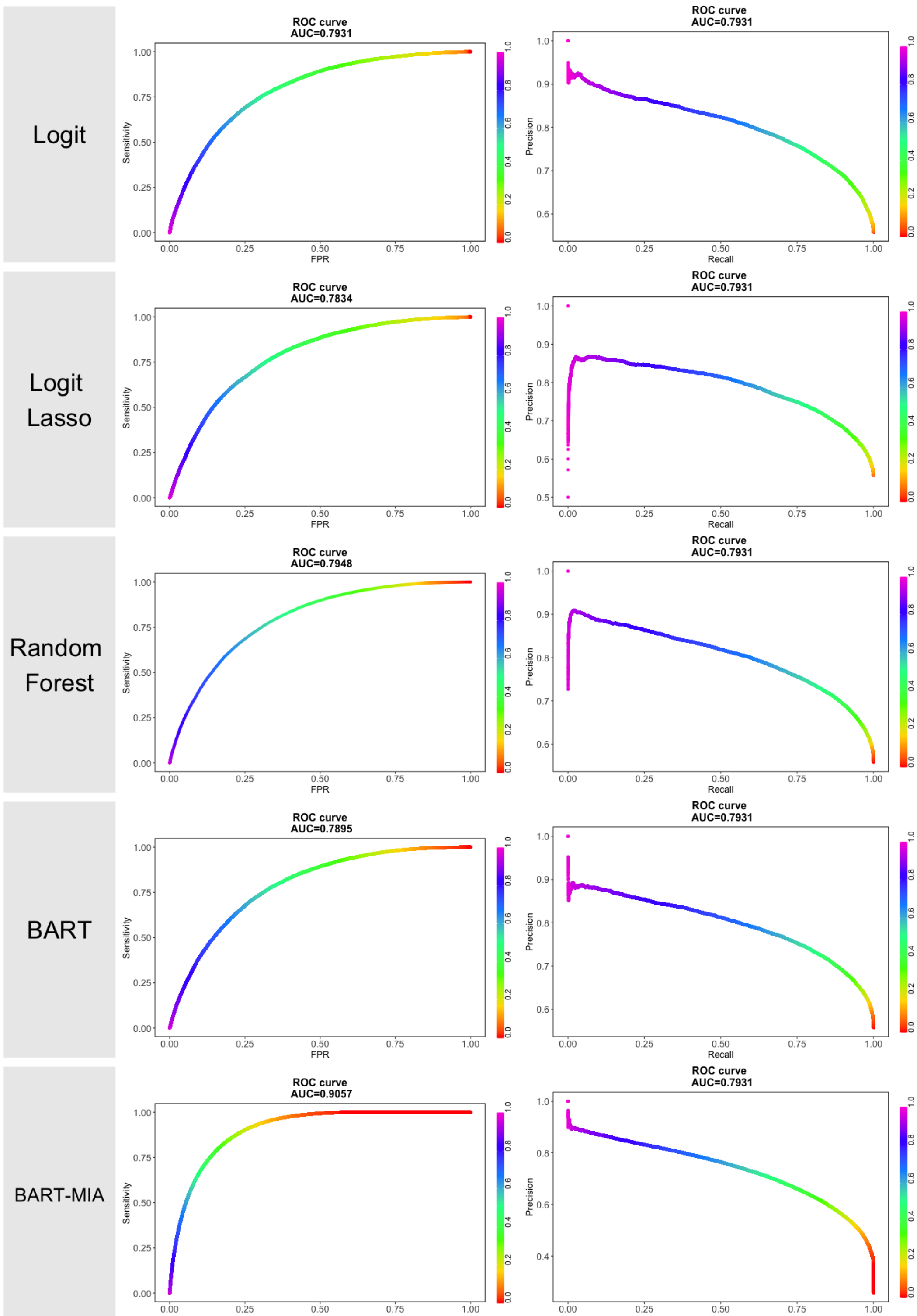
Note: French manufacturing firms are sourced from Orbis, by Bureau Van Dijk. On columns 3 and 4, we separate exporters and non-exporters in our sample. On column 5 we report the total number of manufacturing firms by NACE rev.2. On columns 7-9 a comparison with Eurostat census. When we look at shares on columns 6 and 10, we find our sample is well balanced by industry if compared with the population.

Table B2: Sample coverage - size classes

NACE	Sample - N. employees						Population - N. employees					
rev.2	0-9	10-19	20-49	50-249	250+	Total	0-9	10-19	20-49	50-249	250+	Total
10	1,649	711	611	488	172	3,631	45,798	3,225	1,382	679	204	51,288
11	233	105	93	59	21	511	3,397	205	147	76	28	3,853
13	93	76	107	80	7	363	4,586	209	151	113	17	5,076
14	117	51	49	47	22	286	9,391	140	89	57	16	9,694
15	43	24	36	47	16	166	3,038	70	69	45	21	3,243
16	274	182	178	93	8	735	8,869	560	337	168	21	9,956
17	48	64	105	129	39	385	865	123	121	120	62	1,292
18	381	144	167	86	6	784	14,455	445	277	123	17	15,316
19	1	3	4	6	5	19	NA	NA	3	3	7	25
20	134	109	177	223	87	730	NA	NA	190	219	99	2,515
21	16	18	36	58	61	189	NA	NA	31	50	55	252
22	192	173	274	279	53	971	1,963	405	431	319	86	3,205
23	348	135	161	136	59	839	7,094	266	234	136	72	7,803
24	39	33	53	122	51	298	377	60	56	70	35	599
25	988	792	869	571	75	3,295	13,917	2,174	1,498	734	136	18,460
26	134	113	136	154	70	607	1,700	219	157	171	49	2,295
27	106	83	120	123	64	496	1512	169	168	136	63	2,048
28	281	171	320	319	101	1,192	2,983	455	536	399	160	4,534
29	84	62	103	157	98	504	1,092	156	160	152	75	1,635
30	36	22	30	70	41	199	838	57	63	95	55	1,107
31	148	55	78	66	9	356	8,976	164	134	68	13	9,356
32	311	121	108	102	26	668	20,551	394	217	133	44	21,338
Total	5,656	3,248	3,816	1, 091	3,415	17,226	151,402	9,496,	6,451	4,066	1,335	174,898

Note: French manufacturing firms are sourced from Orbis, by Bureau Van Dijk. Sample coverage by number of employees in 2017 (left panel) is compared with information on population sourced from EUROSTAT Structural Business Statistics. Please note that number of employees may report missing values from sample data, thus number of observations do not sum up to sample totals.

Figure B2: Out-of-sample Goodness-of-Fit



Note: We report the ROC Curves and Precision-Recall curves of the models. See Appendix 11 for the details on the construction of the curves and their interpretation.

Table B3: Prediction accuracies after cross-validating training and testing sets

Measure	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Sensitivity	0.649	0.647	0.654	0.65	0.648
Specificity	0.911	0.904	0.905	0.905	0.907
Balanced Accuracy	0.780	0.775	0.780	0.778	0.778
ROC	0.909	0.903	0.907	0.903	0.908
PR	0.739	0.738	0.742	0.732	0.739
N.Obs	103,540	102,748	102,169	102,028	101,712

Note: We report prediction accuracies of BART-MIA after cross-validating the algorithm on five different random training and testing sets. Our aim is to check whether predictions are robust against data sampling.

Table B4: Prediction accuracies with optimal thresholds (Liu, 2012)

Model	Sensitivity	Specificity	Balanced Accuracy	ROC	PR	Threshold
Logit-Lasso	0.786	0.676	0.716	0.785	0.789	0.513
Logit	0.760	0.688	0.724	0.794	0.805	0.517
Random forest	0.760	0.686	0.723	0.795	0.801	0.560
BART	0.730	0.708	0.719	0.791	0.800	0.569
BART-MIA	0.863	0.791	0.827	0.905	0.738	0.280

Note: We report prediction accuracies when we select the optimal prediction threshold following Liu (2012).

Table B5: Prediction accuracies with a subset of predictors

Model	Sensitivity	Specificity	Balanced Accuracy	ROC	PR
Logit-Lasso	0.668	0.768	0.718	0.786	0.785
CART	0.512	0.907	0.710	-	-
Random forest	0.810	0.627	0.719	0.791	0.793
BART	0.807	0.629	0.718	0.790	0.791
BART-MIA	0.623	0.914	0.768	0.902	0.725

Note: We report prediction accuracies after reducing the battery of predictors from 52 to 23 variables selected by a robust LASSO (Ahrens et al., 2020).

Table B6: Prediction accuracies after training and testing on separate years

Measure	2011	2012	2013	2014	2015	2016	2017	2018
Sensitivity	0.907	0.896	0.885	0.896	0.901	0.918	0.924	0.928
Specificity	0.637	0.632	0.641	0.627	0.639	0.651	0.652	0.654
Balanced Accuracy	0.772	0.764	0.763	0.761	0.770	0.784	0.788	0.791
ROC	0.903	0.889	0.886	0.888	0.894	0.910	0.919	0.930
PR	0.759	0.718	0.725	0.723	0.722	0.729	0.734	0.727
N.Obs	11,375	11,377	11,378	11,383	11,386	11,392	11,388	11,387

Note: We report prediction accuracies of BART-MIA after training and testing on separate years. Our aim is to check whether predictions are robust along the timeline.

Table B7: Prediction accuracies of exporters defined *á la* Békés and Muraközy (2012)

Exporter Class	Sensitivity	Specificity	Balanced Accuracy	ROC	PR	Num. Obs.
Permanent Exporters	0.723	0.779	0.751	0.849	0.934	76,185
Temporary Exporters	0.421	0.820	0.621	0.755	0.447	73,647
Non-Exporters		0.949				158,625
Total	0.650	0.9066	0.7783	0.9048	0.7383	232,272

Note: We report prediction accuracies after BART-MIA for firms classified according to Békés and Muraközy (2012): i) *permanent exporters* are firms that export at least four consecutive years; ii) *temporary exporters* are remaining firms that export at least once; iii) *non-exporters* are firms that never export.

Table B8: Prediction accuracies after an exporters' definition based on thresholds of the share of export revenues over total revenues

Measure	1^{st} Percentile	2^{nd} Percentile	5^{th} Percentile	Benchmark
Sensitivity	0.652	0.641	0.625	0.658
Specificity	0.835	0.837	0.852	0.833
Balanced Accuracy	0.744	0.739	0.738	0.745
ROC	0.836	0.835	0.836	0.836
PR	0.737	0.731	0.724	0.738
N.Obs	41,911	41,911	41,911	41,911

Note: We report prediction accuracies of BART-MIA after defining as exporters the firms with share of export revenues over total revenues above some specific thresholds, at the 1^{st},2^{nd}, and 5^{th} percentiles of the distribution of the share of export revenues over total revenues.

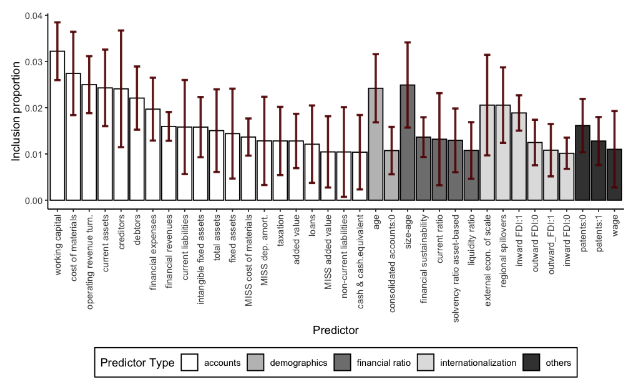
Table B9: Prediction accuracies - Imputation of missing values

	Specificity	Sensitivity	Balanced Accuracy	ROC	PR	N. obs.
LOGIT	0.817	0.751	0.784	0.784	0.528	382,606
LOGIT-LASSO	0.913	0.541	0.727	0.880	0.682	382,606
CART	0.893	0.617	0.755			382,606
Random Forest	0.910	0.647	0.778	0.907	0.738	382,606
BART	0.910	0.635	0.772	0.905	0.731	382,606

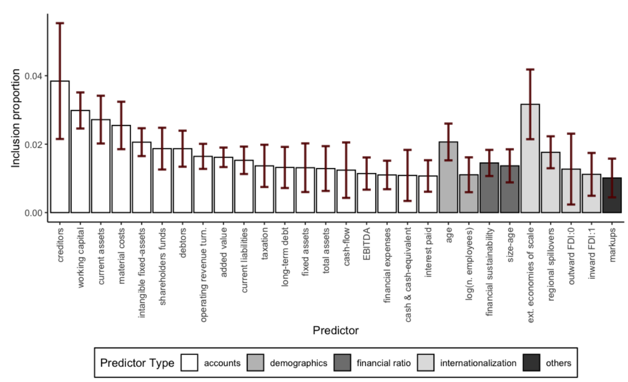
Note: For a robustness check, we report prediction accuracies after an imputation of missing values based on median values, while adding a predictor indicating the number of missing entries by observation (number of missing values by row).

Figure B3: Variable inclusion proportions in Île-de-France *versus* the rest of France

(a) VIPs Île-de-France

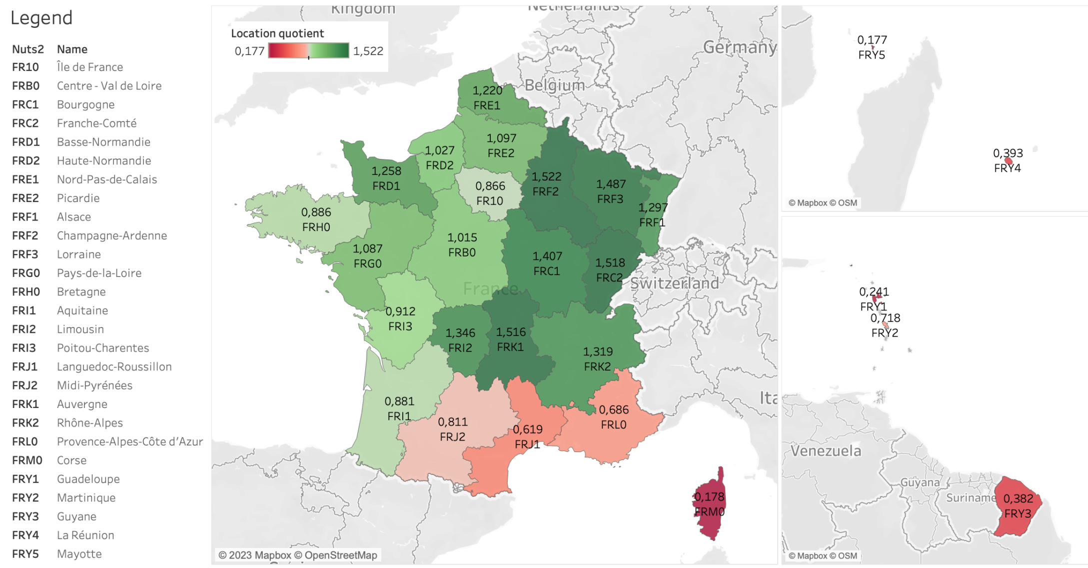


(b) VIPs France, excluding Île-de-France



Note: We report Variable Inclusion Proportions (VIPs) in (a)Île-de-France, (b) in all France *excluding*Île-de-France. Of all the predictors in baseline, we visualize only those with a VIP higher than 1%. The bars represent standard deviations obtained by replicating five different times the BART-MIA on the same random training set.

Figure B4: The potential for extensive margin across France



Note: We report location quotients of non-exporters whose score is above the median in the national distribution. Regions with location quotients greater than one (lower than one) are those where potential exporters are more (less) concentrated than what one would expect given manufacturing density. See Appendix D for details on the computation of location quotients.

Appendix C: Evaluation of prediction accuracy

Different metrics are used to evaluate the prediction accuracy of machine learning algorithms. Briefly, prediction accuracy metrics compare the classes predicted by the algorithm with the actual ones. In the case of a binary outcome, the comparison generates four classes of results:

- **True Positives:** cases when the actual class of the data point is 1 (Positive) and the predicted is also 1 (Positive);
- **False Positives:** cases when the actual class of the data point is 0 (Negative) and the predicted is 1 (Positive);
- **False Negatives:** cases when the actual class of the data point is 1 (Positive) and the predicted is 0 (Negative);
- **True Negatives:** cases when the actual class of the data point is 0 (Negative) and the predicted is also 0 (Negative);

In an ideal scenario, we want to minimize the number of False Positives and False Negatives.

Table B1: Confusion Matrix

		Actual	
		Positives (1)	*Negatives* (0)
Predicted	*Positives* (1)	True Positives (TP)	False Positives (FP)
	Negatives (0)	False Negatives (FN)	True Negatives (TN)

The metrics we use to evaluate prediction accuracy in our exercises are based on the relationship between the sizes of the above classes.

Sensitivity (or Recall) Sensitivity (or Recall) is a measure of the proportion of correctly Predicted Positives out of the total Actual Positives.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Specificity Specificity is a measure that catches the proportion of correctly Predicted Negatives, out of total Actual Negatives.

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

Balanced Accuracy (BACC) Balanced Accuracy (BACC) is a combination of Sensitivity and Specificity. It is particularly useful when classes are imbalanced, i.e., when a class appears much more often than the other. It is computed as the average between the True Positives rate and True Negatives rate.

$$BACC = \frac{Sensitivity + Specificity}{2}$$

Receiving Operating Characteristics (ROC) The ROC curve is a graph showing the performance in classification at different thresholds, expressed in terms of the relationship between True Positive Rate (TPR) and False Positive Rate (FPR), defined as follows:

$$True\ Positive\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$FalsePositiveRate = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

The Area Under the Curve (AUC) of ROC is then useful to evaluate performance in a bounded range between 0 and 1, where 0 indicates complete misclassification, 0.5 corresponds to an uninformative classifier, and 1 indicates perfect prediction.

Precision-Recall (PR) The PR curve is a graph showing the trade-off between Precision and Recall at different thresholds. Note that Precision and Recall are defined as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

As for the ROC curve, the PR AUC is used to evaluate the classifier performance. A High AUC represents both high recall and high precision, thus meaning the classifier is returning accurate results (high precision), as well as returning a majority of all the positive results (high recall).

Appendix D: Location Quotients

Let us define $\mathcal{I} = \{1, \ldots, n\}$ the set of non-exporting firms and $\mathcal{R} = \{1, \ldots, r\}$ the set of regions (NUTS 2-digit). The r partitions of $\mathcal{I}$ by region $j \in \mathcal{R}$ are defined as:

$$I_j \subset \mathcal{I}, j = 1, \ldots, r \quad s.t. \quad \bigcup_{j=1}^{r} I_j = \mathcal{I}$$

Let $\mathcal{P}$ be the set of non-exporting firms whose exporting score e is above the one of the median firm in the total distribution of non-exporters, i.e.:

$$\mathcal{P} \subset \mathcal{I} = \{i \in \mathcal{I} : e_i > median(e)\}$$

Again we can define the r partitions of $\mathcal{P}$ by region $j \in \mathcal{R}$ as

$$P_j \subset \mathcal{P}, j = 1, \ldots, r \quad s.t. \quad \bigcup_{j=1}^{r} P_j = \mathcal{P}$$

The location quotient, for each region $j = 1, \ldots, r$ is computed as

$$LQ_j = \frac{\#P_j / \#I_j}{\#\mathcal{P} / \#\mathcal{I}} \tag{8}$$

In our case, location quotients (LQ) detect the concentration of potential exporters in excess of what one would expect from the national distribution. If, for example, region j has $LQ_j = 1.5$, it implies that firms with a high trade potential are 1.5 times more concentrated in such a region than the average.