Universidad de
SanAndrés

Big data

Profesor: Noelia Romero
Ayudante: Victoria Oubiña

Propuesta de Investigación

Apezteguia, Mannarino, Navajas Jauregui

Fecha: 03/12/2023

Predicción y Aprendizaje de Firmas Exportadoras: Un Estudio para Colombia

Introducción

El crecimiento económico de un país está intrínsecamente vinculado a su capacidad para ampliar sus fronteras comerciales, exportando bienes y servicios a nivel internacional. Sin embargo, esta tarea se presenta como un desafío arduo, dado que implica enfrentarse a una competencia feroz en el mercado global. Para alcanzar el éxito en la exportación, las empresas no solo deben demostrar una eficiencia destacada en sus procesos de producción, sino también en la calidad de sus productos, de manera que puedan satisfacer las demandas internacionales.

A lo largo del tiempo, se ha buscado comprender las complejidades de las transacciones internacionales a través de diversos modelos teóricos. Sin embargo, ha habido una escasa exploración práctica enfocada en la identificación temprana de las empresas con potencial exportador. En este contexto, nuestro objetivo es prever qué empresas se convertirán en exportadoras en Colombia, basándonos en el aprendizaje de aquellas que ya han logrado establecerse en el mercado internacional.

La posibilidad de dar respuesta a este interrogante se materializa gracias a los avances tecnológicos y a la disponibilidad de grandes volúmenes de datos, los cuales han abierto nuevas oportunidades para el desarrollo de modelos de machine learning capaces de analizar, comprender y prever variables de interés.

Contribuir en este ámbito no solo se erige como un elemento clave para la formulación de políticas destinadas a impulsar la exportación, sino también para la asignación eficiente de recursos, el otorgamiento de créditos, capacitaciones sobre el funcionamiento de los mercados internacionales, entre otros aspectos estratégicos. Un modelo de machine learning robusto se convierte así en una herramienta esencial para identificar aquellas empresas que requieren atención especial, con el propósito de impulsar sus actividades de exportación y superar los desafíos iniciales asociados a la adaptación a diferentes entornos regulatorios y de competencia. Profundizando aún más en la propuesta, este análisis nos permitirá identificar las dimensiones críticas que distinguen a las empresas exportadoras de las no exportadoras. Asimismo, posibilitará examinar el comportamiento específico de las exportadoras dentro de cada sector, destacando las similitudes y diferencias

entre aquellas que exportan y las que no, con el objetivo de discernir patrones significativos que guíen estrategias específicas para potenciar el éxito exportador.

El modelo no solo facilitaría la predicción de las firmas exportadoras, sino que posibilitaría identificar variables relevantes que podamos contrastarlas con la teoría existente, con especial hincapié en el modelo teórico de Melitz (2003). Nuestra intención es explorar las variables que, según la teoría, se consideran relevantes para la exportación, y además, buscamos identificar cualquier variable que resulte relevante en el modelo predictivo, pero que no esté contemplada en los marcos teóricos convencionales.

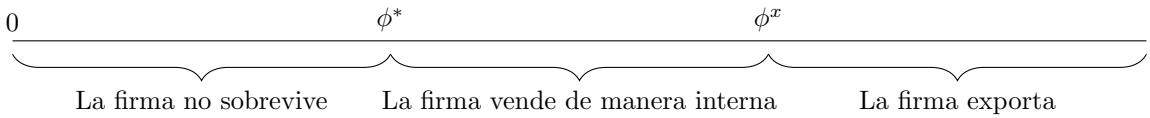
Modelo teórico y literatura previa

El modelo de Melitz (2003) inauguró toda una literatura sobre los determinantes por los cuales una firma es capaz de exportar. En resumidas cuentas, este es el primero en introducir a la firma como la unidad de análisis en la literatura del comercio internacional. Toda la familia de modelos dentro de esta literatura, como el de Dornbusch et. al (1977) ligado a la tradición ricardeana o los modelos ligados a la tradición de Heckscher Ohlin de Deardoff (1979, 1982), no generaban predicciones sobre el tipo de firma que sería capaz de exportar, sino que focalizaban su análisis en los sectores dentro de un país que lo harían. En modelos de economías de escala, por otro lado, todas las empresas son capaces de exportar (Krugman, 1980).

En cuanto al trabajo de Melitz, sobre el cual se hará hincapié en este trabajo, lo que lo hace particularmente revolucionario respecto a la literatura previa es que permite la diferenciación entre empresas. En particular, la variable que diferencia a las empresas es la productividad. Teniendo en cuenta esto, la función de producción de cada empresa será:

$$l = f + \frac{q}{\phi_i}$$

Donde l es la cantidad de trabajo necesaria para producir una unidad q. f es un costo fijo, y $\frac{q}{\phi}$ es un costo variable compuesto por las cantidades a producir y un parámetro de productividad que es particular de cada firma. Cuanto mayor sea la productividad de la firma, menor será su costo unitario medido en trabajo. Esta productividad es una realización de una variable aleatoria que se devela luego de pagar un costo de entrada f_e. Teniendo en cuenta la existencia de costos fijos, existirá una productividad umbral (ϕ^*) que permite a la empresa individual seguir operando obteniendo algún tipo de ganancia. En esta misma línea, para poder exportar, se necesitará una productividad (ϕ^x) tan alta capaz de costear el costo fijo de hacerlo ($f_x > f$); en este caso se están modelando aquellos costos que hay que pagar para poder exportar, como el ingreso al mercado, ciertos estándares sanitarios en algunos casos, etc. En resumen, para aquellas firmas que pagan el costo de entrada, su nivel productividad las hará mantenerse en el mercado, salir o les dará la posibilidad de exportar:



El principal hecho estilizado que debería observarse en la realidad, si el modelo es correcto, es que las firmas de mayor tamaño son capaces de exportar. Sin embargo, varios trabajos han encontrado que esto no se cumple y propusieron diversas extensiones. Entre los estudios relevantes en esta área se incluyen los trabajos de Arkolalis (2010), Kugler y Verhoogen (2012), y Hallak et al. (2013), los cuales extienden el modelo seminal de forma tal que sea capaz de explicar cuestiones de la realidad para las cuales originalmente no tenía respuestas.

En Hallak et al. (2013) se hace hincapié en la calidad como otro posible factor de la heterogeneidad de las empresas. Los mercados internacionales requieren además de un precio competitivo, el cual puede explicarse según la eficiencia del proceso productivo, un margen de calidad mayor al demandado localmente y acorde a los estándares internacionales. La calidad, en este sentido, puede servir como un ajuste a la productividad ϕ. Esta posible extensión explica por qué empresas más grandes, con procesos productivos eficientes, no son capaces de exportar: su calidad no es lo suficientemente alta para ser valorada en mercados externos.

Otra predicción del modelo es que aquellas empresas que logran pasar el umbral de productividad para poder exportar lo harán a todos los mercados existentes, en otras palabras, exportarán a todos los países. Sin embargo, este hecho estilizado no se observa en la realidad. En el trabajo de Arkolakis (2010) se extiende el modelo agregando una función de costo creciente en los mercados que se quieran alcanzar. Esta nueva configuración es acorde con el hecho que las empresas más productivas del subconjunto de las exportadoras llegan a una menor cantidad de mercados.

Nuestro trabajo buscará hacer un aporte a esta literatura, intentando identificar aquellas variables que son relevantes a la hora de definir el futuro exportador de una firma. Lo novedoso es el enfoque empírico, en el sentido de utilizar firmas que exportan como fuente de identificación de sus características principales. Nuestro trabajo se enfoca en el segmento antes presentado donde las firmas logran sobrevivir y el *cutoff* de las que logran exportar. Buscamos así encontrar firmas que estén cerca del umbral, las cuales son plausiblemente las más adecuadas para enfocar la política productiva tendiente al fomento de las exportaciones.

Además, el trabajo se enmarcará en la literatura relacionada con Machine Learning. Un paper cercano a este trabajo es el de Micocci y Rungi (2022)[1]. Los autores realizan un modelo predictivo para Francia en el período de 2010 a 2018, centrado en estados financieros, siguiendo la lógica utilizada por las instituciones financieras para predecir el riesgo crediticio. Nuestro trabajo, por otro lado, cuenta con una variedad más importante de variables relevantes sobre las cuales realizar la predicción. Por otro lado, en el artículo de Jia, Adland y Wang (2021) se emplean métodos de machine learning para anticipar el destino de las exportaciones

[1]La última versión que encontramos es la de septiembre de 2022

de petróleo, aprovechando datos detallados de envíos de petróleo a nivel micro. Creemos que nuestro trabajo puede hacer un gran aporte, especialmente debido a la riqueza de la base de datos con la que contamos para lograr nuestro objetivo.

Base de Datos

El objetivo es realizar el modelo para Colombia, ya que dispone de una base de datos amplia y continua en el tiempo. El Departamento Administrativo Nacional de Estadística (DANE)[2] ofrece dos encuestas de particular interés para nuestro estudio. Estas encuestas se centran en los establecimientos con 10 o más personas ocupadas o que en su defecto registren un valor de producción anual igual o superior a un valor que se especifica para cada año de referencia informado por el directorio de las empresas.

Nuestra primera base de datos es la Encuesta Anual Manufacturera (EAM), que se realiza desde 1992 y abarca más de 300 variables. Entre ellas se encuentran la identificación o ID de la empresa, la ubicación (región), CIIU (actividad industrial), el personal ocupado, salarios, consumo intermedio, valor agregado, inversiones, energía consumida, cantidad exportada, entre otras. El objetivo de esta encuesta es poder conocer la estructura de la empresa, medir la evolución y comportamiento del sector industrial, determinar la composición de la industria nacional y obtener la distribución regional de la actividad industrial.

Nuestra segunda base de datos es la Encuesta de Desarrollo e Innovación Tecnológica en la Industria Manufacturera (EDIT), la cual se lleva a cabo desde 2007 y comprende más de 400 variables, que en su mayoría son variables dummies. Estas variables incluyen nuevamente la identificación, introducción de nuevos bienes y servicios en la empresa/mercado, cantidad de innovaciones, expansión de mercado, cambios en insumos, escasez de información, préstamos, participación en programas nacionales, calidad de empleados, entre otras. Esta encuesta busca acercarse a la empresa desde otra óptica, buscando conocer la forma en que se desarrolla. En este contexto, busca recabar información estadística que dé cuenta acerca del avance de la innovación y el desarrollo tecnológico. Esta información es herramienta fundamental en la generación y uso del conocimiento a favor del aprovechamiento de oportunidades empresariales y la introducción de nuevos productos y servicios al mercado.

Dado que ambas encuestas comparten un identificador único de la firma (nordemp), podemos combinarlas para tener una única base de datos, que comenzará a tener registros desde el año 2007, cuando se inició la encuesta EDIT. Esto nos proporcionará un panel con más de 5000 observaciones. Debemos agregar variables dummies relacionadas con acuerdos comerciales por sector entre países, así como una variable dummy para cada año que indique si la empresa exporta o no.

A continuación, se presentan tablas descriptivas correspondientes a las empresas exportadoras y no ex-

[2]El link a la base de datos se encuentra Aquí

portadoras del año 2014, basadas en la Encuesta Anual Manufacturera (EAM). Estas tablas contienen información relevante para comprender la estructura de las empresas

La base de datos inicial se compone de 1.795 empresas exportadoras y 7.451 empresas no exportadoras. Al analizar las estadísticas, observamos que la media de la producción industrial en las empresas exportadoras es aproximadamente cinco veces mayor que en las no exportadoras. Asimismo, la composición de activos fijos y el consumo de energía eléctrica son significativamente mayores en las empresas exportadoras en comparación con las no exportadoras. En cuanto al personal total, las empresas exportadoras presentan, en promedio, el triple de empleados en comparación con las no exportadoras.

Cuadro 1: Empresas exportadoras: 1795

	Producción industrial (Mill.)	Activos fijos (Mill.)	Energía eléctrica en Kw	Personal total
Min	0.15	0	0	1
Max	6721.94	2321.69	245.78	2869
Mean	68.69	45.67	5.88	195.13

Cuadro 2: Empresas no exportadoras: 7451

	Producción industrial (Mill.)	Activos fijos (Mill.)	Energía eléctrica en Kw	Personal total
Min	0	0	0	1
Max	1028.59	1794.04	417.51	1590
Mean	11.92	5.76	0.7	47.53

Es importante destacar que, aunque generalmente se espera que las empresas de mayor tamaño sean más propensas a la exportación, encontramos casos atípicos dentro de las empresas no exportadoras, donde algunas tienen hasta 1590 empleados, superando considerablemente la media de empleados en las empresas exportadoras y no son empresas exportadoras. Esto va en línea con algunos de los papers mencionados en la sección anterior.

Metodología

En rasgos generales, la metodología propuesta será utilizar un modelo predictivo de Machine Learning de manera tal que, a través de la metodología de Elastic Net, se determinen las variables relevantes según los datos de nuestra muestra de entrenamiento para poder predecir si la empresa puede exportar o no.

En los datos de nuestra muestra podemos observar si las empresas exportaron o no, por lo que podemos utilizar técnicas de aprendizaje supervisado. El principal desafío que enfrentamos en este proceso es concluir si hay linealidad o no en nuestros datos, es decir, si las variables son relevantes siempre o condicional al

resultado de otras variables.

En el caso de que existiese linealidad en los datos, la estrategia será realizar la regularización utilizando Naive Elastic Net, ya que notamos que en nuestros datos existe alta correlación entre las variables, y en este escenario esta técnica funciona mejor que Lasso. Utilizaríamos Naive Elastic Net y no Elastic Net, ya que este último genera demasiado sesgo al encoger dos veces los coeficientes y funciona peor en la práctica. Luego, para realizar la clasificación, descartamos la técnica de vecinos más cercanos, ya que presenta problemas cuando las muestras se encuentran no balanceadas, y sospechamos que este es un problema de nuestra muestra. De cualquier manera, teniendo en cuenta la naturaleza de nuestros datos, decidimos profundizar la metodología para datos no lineales.

Si, tal como sospechamos, nuestros datos no son lineales, consideramos conveniente utilizar la estructura de árboles de decisión para reducir la dimensionalidad de nuestro modelo, particularmente utilizando Weakest Link Prunning. Luego, para realizar la clasificación decidimos que lo más adecuado a nuestro objetivo es utilizar Random Forest. En un primer momento, consideramos también la opción de Bagging. Sin embargo, a pesar de que Random Forest toma un subset de variables aleatorias más chicos que Bagging, aleatorizar las variables que entran en este subset genera que exista menos correlación entre los árboles, por lo que en nuestro escenario de alta correlación entre variables, esta metodología es más apta, ya que permite que obtengamos resultados que son más robustos. Teniendo en cuenta que nuestro objetivo es ver si una empresa va a exportar o no y que una posible aplicación será utilizar esto para hacer *policy* sobre las variables más importantes, decidimos que lo más conveniente era utilizar Random Forest por sobre Bagging. Consideramos en este proceso también otras posibilidades como ANN (redes neuronales artificiales), pero las descartamos debido a la dificultad que representa para interpretar resultados.

Para poder decidir cuál de estas dos alternativas seguir, no pudimos utilizar justificaciones teóricas de la literatura existente, ya que no encontramos discusiones sobre la linealidad o no de datos en el contexto del carácter de una firma como exportadora o no. Frente a esto, debimos realizar un análisis de nuestros datos para poder tomar una decisión acerca de qué enfoque íbamos a realizar la predicción. Para esto, en primer lugar realizamos gráficos de dispersión para poder observar la relación entre las variables. Aquí, nos llamó la atención que los datos se agrupaban formando curvas, principalmente cuando graficamos la relación del tamaño de la firma con demás variables. Asimismo, observamos agrupaciones de datos que nos llamaron la atención, sirviendo como los primeros indicios de no linealidad.

Además, para no quedarnos únicamente con un criterio de decisión, calculamos los errores de medición utilizando modelos para datos lineales y no lineales, y obtuvimos mejores resultados con los modelos de ensamble de árboles. Aunque este análisis es preliminar, lo consideramos suficiente para concluir que en nuestros datos existe no linealidad. Además, esto coincide con lo que nosotros considerábamos antes de realizar estos análisis, ya que al hacer un análisis meramente descriptivo observando la base de datos, notábamos que existían variables que se volvían relevantes, condicional al valor de otra variable. Particularmente, esto

lo notamos al analizar la relación del tamaño de la firma con las demás variables. La teoría de comercio internacional indica que el tamaño de la firma es una variable relevante a la hora de decidir si una firma exporta o no, y al analizar nuestra base podíamos ver que las empresas más grandes eran más propensas a exportar. Sin embargo, notamos que otras variables que eran pequeñas en firmas exportadoras grandes, tomaban valores altos en casos en los que el tamaño de la firma era pequeño, y esas firmas, a pesar de su tamaño, exportaban. Consideramos que esto puede estar explicado justamente por la idea de que cuando la firma es pequeña, hay otras variables que toman relevancia a la hora de determinar si la firma exporta o no, y esto se corresponde con la existencia de datos no lineales.

De esta manera, combinando el análisis técnico con un análisis más descriptivo, concluimos que nuestros datos son no lineales, por lo que la metodología que decidimos utilizar para realizar nuestra predicción es las que describimos al hablar de no linealidades.

Comentarios finales

El objetivo de este trabajo fue desarrollar un modelo con un alto nivel de precisión que pueda predecir de manera efectiva qué empresas se convertirán en exportadoras. En este sentido, como un subproducto de esta práctica y, teniendo en cuenta la riqueza de nuestros datos, también podemos ser capaces de identificar las características principales de aquellas empresas que exportan. Debido a esto, un aspecto relevante será el diálogo entre la teoría y la práctica, ya que nuestro modelo nos permitirá identificar determinantes de la exportación empresarial. Las conclusiones de nuestro modelo servirán como un insumo de la política pública, pudiendo focalizar los esfuerzos en aquellas empresas que tengan ciertas características que las coloquen cerca del umbral que la teoría predice debe superarse para exportar. Es en este sentido que esperamos poder realizar una contribución en el campo de desarrollo económico y el comercio internacional, con la puesta en práctica de herramientas modernas para brindar una nueva opción a utilizar al momento de la elección de las políticas productivas.

Por otro lado, la determinación de las características que comparten aquellas firmas que son capaces de exportar servirán como evidencia adicional de los modelos teóricos antes citados. Teniendo en cuenta los datos analizados de manera preliminar, el tamaño de la empresa, un aspecto ampliamente resaltado por la literatura, parece servir como buen predictor de la capacidad exportadora de la firma. Esperamos, además, encontrar características adicionales que no hayan sido remarcadas previamente en la literatura y hacer también un aporte en este sentido.

Por último, aunque no es el objetivo de este trabajo, la riqueza de nuestra base de datos nos permitiría realizar análisis causales a través de diseños de regresión discontinua.

Bibliografía

- Arkolakis, C. (2010). Market penetration costs and the new consumers margin in international trade. *Journal of political economy, 118(6)*, 1151-1199.
- Deardorff, A. V. (1979). Weak links in the chain of comparative advantage. *Journal of International Economics, 9(2)*, 197-209.
- Deardorff, A. V. (1982). The general validity of the Heckscher-Ohlin theorem. *The American Economic Review, 72(4)*, 683-694.
- Dornbusch, R., Fischer, S., & Samuelson, P. A. (1977). Comparative advantage, trade, and payments in a Ricardian model with a continuum of goods. *The American Economic Review, 67(5)*, 823-839.
- Hallak, J. C., & Sivadasan, J. (2013). Product and process productivity: Implications for quality choice and conditional exporter premia. *Journal of International Economics, 91(1)*, 53-67.
- Haiying Jia, Roar Adland, and Yuchen Wang, (2021). Global Oil Export Destination Prediction: A Machine Learning Approach. *The Energy Journal, International Association for Energy Economics, vol. 0(Number 4)*.
- Kugler, M., & Verhoogen, E. (2012). Prices, plant size, and product quality. *The Review of Economic Studies, 79(1)*, 307-339.
- Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica, 71(6)*, 1695-1725.
- Micocci, F., & Rungi, A. (2023). Predicting Exporters with Machine Learning. *World Trade Review, 22(5)*, 584-607.