Universidad de
SanAndrés

E337 - Big data

Propuesta de Investigación

Predicción del crecimiento del PBI de Argentina utilizando Random Forest

Autores:

Rodrigo Braga

Fernando Kricun

Julián Ramos Cancio

Fecha de entrega: 3 de diciembre de 2023

Introducción

El presente trabajo tiene como objetivo presentar una alternativa para la realización de una proyección del crecimiento del PBI real de Argentina utilizando técnicas de *machine learning* o aprendizaje automático, en particular, *Random Forest*.

La predicción de indicadores macroeconómicos utilizando la evolución histórica y temporal de distintas variables proporciona información útil y relevante para la toma de decisiones de los distintos agentes económicos, ya sea para los hacedores de políticas públicas, las empresas o los hogares. Estas predicciones se basan en el análisis de datos en series temporales, como el PBI real, la tasa de inflación, la tasa de desempleo, entre otros. En este contexto, donde hay evidentes dependencias temporales, el enfoque estándar de muestras independientes carece de sentido. Por ello, para realizar sus pronósticos los métodos tradicionales recurren a modelos autorregresivos en los cuales las variables dependen de sus propios rezagos. La elección de estos rezagos no es una cuestión meramente arbitraria y requiere realizar ciertos supuestos en los modelos.

Asimismo, si deseamos pronosticar el PBI real de un país específico, es necesario incorporar esta variable como dependiente en relación con diversas variables independientes, algunas de las cuales pueden incluir rezagos. Este procedimiento implica un proceso que requiere intuición económica y la formulación de varios supuestos por parte del analista que realiza la proyección. Cualquier error en estas suposiciones podría conducir a predicciones imprecisas o incorrectas por parte de los modelos.

En este sentido y en contraste con los modelos de pronóstico tradicionales, los modelos de aprendizaje automático trabajan casi en su totalidad con pura predicción (Varian, 2014). Estos modelos son más flexibles que los tradicionales y pueden producir predicciones sin la necesidad de realizar tantos supuestos. Además, por lo general, suelen predecir con mayor precisión e incluso son útiles cuando se cuenta con bases de datos de baja frecuencia (Yoon, 2020).

En la literatura, se ha explorado el desarrollo de diversas técnicas de *machine learning*, y sus resultados en términos de predicciones han sido más que satisfactorios. Sin embargo, en el caso específico de Argentina, parece que esta cuestión no ha sido suficientemente investigada, y solo se encuentran algunos trabajos al respecto. Es precisamente en este contexto donde radica el valor de nuestro estudio. Dada la delicada situación económica que atraviesa el país, cualquier grado de certidumbre en relación con los indicadores macroeconómicos futuros adquiere una importancia aún mayor que en la mayoría de los países del mundo.

Recientemente, el presidente electo Javier Milei anticipó que en los próximos meses Argentina atravesará un período de estanflación, entendido como un período de alta inflación que coincide con una caída de la actividad económica. En un contexto de semejantes características, es necesario contar con los métodos de predicción más precisos posibles, de modo tal que los agentes económicos cuenten con la mejor información posible a la hora de tomar sus decisiones.

La presente propuesta estará organizada de la siguiente manera: en la próxima sección haremos una breve revisión de la literatura que hace uso de técnicas de machine learning

para realizar pronósticos de distintas variables económicas. En mayor parte, mencionaremos trabajos que emplean distintos métodos para varios países del mundo. Posteriormente, analizaremos distintas bases de datos que se pueden utilizar para realizar un pronóstico del crecimiento del PBI real en Argentina. Luego, desarrollaremos la metodología a utilizar; en particular, nuestras predicciones se basarán en el método de random forest. Finalmente, brindaremos unas breves conclusiones con los resultados que esperamos encontrar si se llevase a cabo esta propuesta.

La estructura de la presente propuesta se organizará de la siguiente manera: en la próxima sección, se realizará una breve revisión de la literatura que emplea técnicas de aprendizaje automático para pronosticar diversas variables económicas. En su mayoría, se destacarán trabajos que utilizan diversos métodos en diferentes países del mundo. Posteriormente, se llevará a cabo un análisis de diversas bases de datos que podrían ser utilizadas para prever el crecimiento del PBI real en Argentina. Luego, se describirá la metodología a emplear, centrándonos específicamente en el método de *random forest*. Finalmente, se ofrecerá una breve conclusión con respecto a los resultados esperados en caso de llevar adelante esta propuesta.

Literatura previa

La mayoría de los estudios sobre este tópico parece indicar que los modelos de predicción que utilizan *machine learning* son más precisos que los modelos autorregresivos de series de tiempo tradicionales.

Paruchuri (2021) replica una predicción del PBI real de Italia desde el primer trimestre del año 1995 hasta el segundo trimestre del año 2015. Sus resultados indican que los algoritmos de aprendizaje automático logran un rendimiento superior en la predicción de las recesiones económicas que los métodos estadísticos convencionales.

Siguiendo esa línea, el estudio de Yoon (2020) realiza una predicción del crecimiento del PBI real de Japón para el período comprendido entre los años 2001 y 2018 mediante algoritmos de *machine learning*. En particular, Yoon realiza las predicciones utilizando los métodos de gradiente y de *random forest*. Luego, compara sus proyecciones con las tradicionales que realiza el Banco de Japón y el Fondo Monetario Internacional (FMI). Sus resultados coinciden con los de Paruchuri; las proyecciones basadas en *machine learning* son más precisas que aquellas realizadas por los organismos mencionados.

Martin (2019) también realiza una comparación entre los métodos de aprendizaje automático y los modelos de proyección autorregresivos tradicionales con el objetivo de evaluar el rendimiento de estas técnicas. Ambos enfoques se utilizan para proyectar el cambio porcentual del PBI de Sudáfrica, trimestre a trimestre. Los resultados indican que los métodos computacionales superan a los convencionales, según criterios de minimización de la raíz del error cuadrático medio y maximización de la correlación con la tendencia actual de los datos. Con estos hallazgos, la autora sugiere que estas metodologías innovadoras representan una opción viable para proporcionar información sobre la tendencia de diversas variables macroeconómicas a los responsables de la formulación de políticas públicas, mejorando así el proceso de toma de decisiones mediante datos más

precisos.

Richardson, van Florenstein Mulder y Vehbi (2018) también llevan a cabo una comparación, en esta ocasión, trabajando con el PBI de Nueva Zelanda. En su estudio, los autores contrastan el rendimiento de los métodos computacionales con otras técnicas tradicionales, como la autoregresión vectorial bayesiana. Llegan a la conclusión de que, en la mayoría de los casos, los modelos de *machine learning* tienen la capacidad de generar pronósticos más precisos en comparación con los modelos autorregresivos y otros enfoques estadísticos.

Sin embargo, existen otras investigaciones que llegan a resultados opuestos, es decir, muestran que los modelos tradicionales son más precisos que los basados en aprendizaje automático. En este sentido, Kurihara y Fukushima (2019) llevaron a cabo una comparación entre las proyecciones realizadas mediante *machine learning* y las efectuadas con modelos autorregresivos para el PBI y los precios al consumidor en los países del G7. Los resultados empíricos indican que, en este caso particular, los pronósticos tradicionales son más precisos que los modelos computacionales. No obstante, la diferencia entre ambas técnicas al pronosticar los precios al consumidor es mínima.

Por otra parte, Premraj (2019) realiza una comparación entre los métodos analizados en 10 países: Australia, Canadá, Euro Área, Alemania, España, Francia, Japón, Suecia, Gran Bretaña y Estados Unidos. Sus resultados muestran que los modelos de regresión tradicionales de series de tiempo son superiores a aquellos basados en *machine learning* en la gran mayoría de los casos. El autor sugiere que este resultado puede deberse a que aún no existen suficientes variables macroeconómicas y observaciones para que el algoritmo pueda aprender de manera efectiva. Además, destaca que en el caso de los Estados Unidos, donde hay una gran cantidad de información disponible, los algoritmos de aprendizaje automático proporcionaron pronósticos más precisos.

Chu y Qureshi (2022) realizan una comparación entre las proyecciones del PBI de Estados Unidos utilizando técnicas de machine learning y de deep learning y llegan a que, en la mayoría de los casos, los métodos de machine learning son mejores y más precisos que los basados en deep learning.

Es también relevante destacar algunas investigaciones que emplean algoritmos de aprendizaje profundo (*deep learning*), un subgrupo del *machine learning* que enfatiza el aprendizaje de capas sucesivas de representaciones cada vez más significativas (Alegre y Lozano, 2022). El estudio de Alegre y Lozano (2022) utiliza diversas técnicas de *deep learning* en series de tiempo para realizar proyecciones macroeconómicas en diferentes países de América Latina, incluyendo Argentina. Esta investigación se basa en series macroeconómicas anuales y trimestrales extraídas del Banco Mundial y de la Comisión Económica para América Latina y el Caribe (CEPAL).

Finalmente, Chu y Qureshi (2022) llevan a cabo una comparación entre las proyecciones del PBI de Estados Unidos utilizando técnicas de *machine learning* y *deep learning*, concluyendo que, en la mayoría de los casos, los métodos de aprendizaje automático resultan superiores y más precisos que los basados en aprendizaje profundo.

Base de datos

La base de datos que se propone en este trabajo consta de un conjunto de series para Argentina que reflejen el desempeño del país en materia económica desde distintos sectores. A modo de ejemplo, Yoon (2020) trabaja 23 variables explicativas que brindan información sobre distintos campos de la macroeconomía, a saber: consumo, balanza de pagos, tipo de cambio, entre otras variables.

Para el presente estudio se sugiere el armado de una base de datos con indicadores de PBI, nivel de precios, consumo, importaciones, exportaciones, estado de la balanza de pagos, tasa de empleo y desempleo, tipos de cambio, inversión extranjera directa, tasa de interés, reservas y porcentaje de deuda sobre PBI. Estos indicadores pueden extraerse a partir de distintas fuentes, a saber: Instituto Nacional de Estadística y Censos (INDEC), Banco Mundial, Fondo Monetario Internacional (FMI) y portales de estadística como Statista [1].

Una cuestión importante a señalar es el hecho de que no se propone modificar o corregir las series de la base de datos. Esto es algo que en otros trabajos sí se realiza (Alegre y Lozano, 2022). La justificación detrás de estas correcciones parte de la base de que los métodos de *machine learning* encuentran dificultades a la hora de predecir escenarios condicionados por acontecimientos improbables, y esta es una situación que en los últimos 20 años ocurrió dos veces: crisis del 2008 y la pandemia de COVID-19. Por lo tanto, en aquellos estudios que comparan la precisión de métodos computacionales puede tener sentido realizar estas correcciones. Sin embargo, debido a que el presente trabajo tiene como objetivo la comparación entre metodologías tradicionales y de *machine learning*, corregir las series sobreestimaría la precisión de estas últimas.

Metodología

Para esta sección, hemos decidido replicar el método de Random Forest utilizado por Yoon (2020) en Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach.

Utilizando datos trimestrales del período comprendido entre 1990-2022, el modelo de random forest podría predecir el crecimiento anual del PBI real de Argentina desde años recientes hasta el año 2022. Esto es porque parte de los años inciales se utilizan para entrenar el modelo. Además, el modelo es diseñado para realizar predicciones del crecimiento anual del PBI real de un determinado año basándose en datos hasta el segundo trimestre de ese año inclusive. Por ejemplo, para predecir el año 2010, el modelo entrena y ajusta usando datos hasta el segundo trimestre del 2010. Esto asegura que no se utilice información del futuro para predecir información pasada, como bien señala Yoon (2020). Este método de random forest fue introducido inicialmente por Breiman (2001) y usa árboles de regresión. Utilizando datos bootstrapeados, los árboles de regresión son entrenados independientemente y su resultado es promediado para luego producir las predicciones (Yoon, 2020).

Este método de random forest fue introducido inicialmente por Breiman (2001) y usa árbo-

[1] Sitio web: https://es.statista.com/

les de regresión. Utilizando datos bootstrapeados, los árboles de regresión son entrenados independientemente y su resultado es promediado para luego producir las predicciones (Yoon, 2020).

Los pasos básicos del modelo de random forest son los siguientes:

En primer lugar, se crea un conjunto de muestras bootstrapeadas (llamado M) de tamaño N de los datos de entrenamiento. Luego, para cada una de esas muestras se arma un random forest tree T_m aplicando recursivamente los siguientes pasos:

1. Seleccionar aleatoriamente m variables de los p predictores posibles
2. Tomar la mejor variable y el mejor punto de partición dentro de los m predictores
3. Dividir el nodo en dos nuevos nodos. La división es decidida de forma tal que minimice el error cuadrático medio (MSE), que es calculado de la siguiente forma:

$$F_0(x) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \gamma\right)^2$$

Donde y_i es el valor observado y gamma es el valor esperado.

Por último, con el conjunto de árboles resultantes, se calcula la predicción promediando los resultados de todos los árboles:

$$\hat{F}_{rf}^{M}(x) = \frac{1}{M}\sum_{m=1}^{M} T_m(x).$$

Este algoritmo previamente detallado puede ser implementado utilizando el paquete *Scikit-Learn* de Python.

Por otro lado, para la selección de los hiperparámetros, utlizaremos la técnica de *k-fold cross-validation* (Figura 1), con un k=5 como aconseja la literatura. Este método separa los datos de entrenamiento en k piezas y testea separadamente cada pieza para ajustar el modelo. Además, debido a la dependencia temporal propia de los datos de series de tiempo, la técnica de *k-fold cross-validation* es diseñada para establecer los primeros *k-folds* como el conjunto de entrenamiento y los datos después del fold como el conjunto de prueba o test (Yoon, 2020).

Conclusiones y limitaciones

A la hora de ejecutar esta propuesta, esperamos encontrar resultados similares a los que se obtienen en la mayor parte de la literatura, es decir, esperaríamos encontrar que este método de predicción basado en random forest sea más preciso que los pronósticos autorregresivos tradicionales.

Sin embargo, una de las limitaciones a las que nos podemos enfrentar es que la disponibilidad de los datos para la Argentina no sea suficiente para que el algoritmo logre aprender. En ese hipotético caso, esperaríamos encontrar que los modelos tradicionales tengan un mejor desempeño. Pese a esto, sea cual sea el resultado de esta aplicación, esperamos que esta propuesta sea el puntapié inicial para incentivar un abordaje y desarrollo más profundo de las distintas técnicas de *machine learning* para la proyección de distintas variables macroeconómicas en la Argentina.

Bibliografía

Alegre Ibáñez, V. A., & Lozano Aparicio, J. M. (2022). Aplicación de métodos de *Deep Learning* en series de tiempo para el pronóstico de la situación macroeconómica en América Latina. *Interfases*, 15(015), 102-130. https://doi.org/10.26439/interfases2022.n015.5817

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32

Chu, B., & Qureshi, S. (2022). Comparing Out-of-Sample Performance of Machine Learning Methods to Forecast U.S. GDP Growth. *Computational economics*, 1–43. Advance online publication. https://doi.org/10.1007/s10614-022-10312-z

Kurihara, Y., & Fukushima, A. (2019). AR Model or Machine Learning for Forecasting GDP and Consumer Price for G7 Countries. *Applied Economics and Finance*, 6(3), 1-10. https://doi.org/10.11114/aef.v6i3.4126

Martin, L. (2019). Machine Learning vs Traditional Forecasting Methods: An Application to South African GDP. Working Paper from Stellenbosch University, Department of Economics (No. 12/2019).

Paruchuri, H. (2021). Conceptualization of Machine Learning in Economic Forecasting. *Asian business review*, 11, 51-58. https://doi.org/10.18034/abr.v11i2.532

Premraj, P. (2019). Forecasting GDP growth: a comprehensive comparison of employing machine learning algorithms and time series regression models.

Richardson, A., van Florenstein Mulder, T., Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, 37(2), 941-948. https://doi.org/10.1016/j.ijforecast.2020.10.005

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.

Yoon, J. (2021). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. *Computational Economics*, 57(2), 247–265. https://doi.org/10.1007/s10614-020-10054-w

Anexo

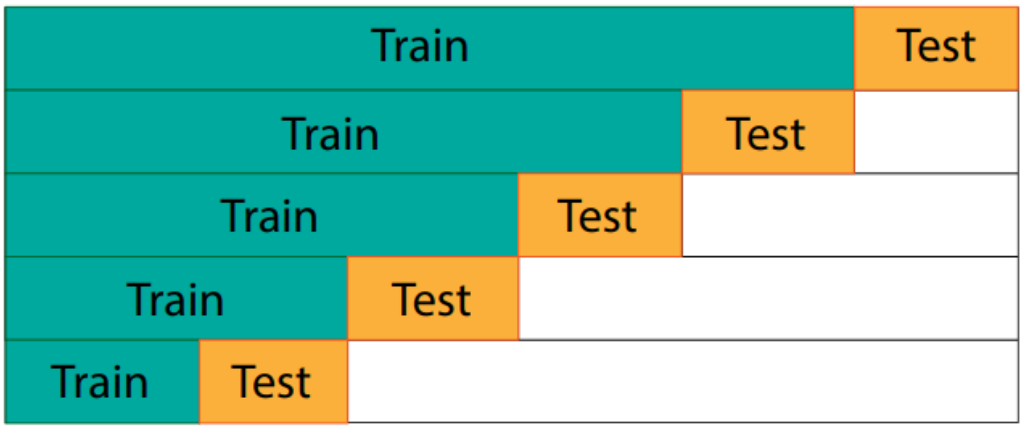

Figura 1: Proceso de cross-validation