

PulsoPBA

El termómetro de actividad de la Provincia de Buenos Aires



Metodología PulsoPBA

Introducción

PulsoPBA es una herramienta creada para mostrar la evolución de la actividad económica de la Provincia de Buenos Aires (PBA), desarrollada por la Gerencia de Estudios Económicos del Banco Provincia en colaboración con el Instituto de Ciencias de la Computación de la Universidad de Buenos Aires (ICC-UBA).

El Banco posee bases de datos de alta frecuencia provenientes de sus operaciones bancarias. Estas incluyen información detallada sobre transacciones, préstamos, operaciones de comercio exterior y consumos a través de diferentes medios de pago. Los datos están desagregados por fecha, centro zonal, actor institucional, actividad económica (CLANAE a dos dígitos), tipo de moneda y tipo de movimiento. A partir de esta información, se construyó un conjunto de aproximadamente 1.200 variables agregadas con frecuencia diaria, que luego a su vez se agregan en bloques semanales por cada mes.

Utilizando estas bases y el Estimador Mensual de Actividad de la Provincia de Buenos Aires (EMA-PBA), elaborado por el Ministerio de Economía de la Provincia de Buenos Aires, se entrenaen algoritmos de Inteligencia Artificial (IA) diseñados para modelar series de tiempo. Una vez entrenados, los modelos permiten estimar el nivel actual de actividad económica (nowcasting), proyectar su evolución en las semanas siguientes (forecasting) y realizar un reentrenamiento mensual. Tanto las estimaciones como el reentrenamiento se generan de forma automatizada en un servidor de la Gerencia de Estudios Económicos.

Este documento presenta la metodología utilizada para la construcción del sistema de modelos de predicción de PulsoPBA.

Definiciones y conceptos básicos para comprender la metodología

1. EMA-PBA: Estimador mensual de actividad económica de la Provincia de Buenos Aires, serie sin estacionalidad. Indicador oficial mensual que adelante el PIB trimestral. Se publica periódicamente luego de la publicación del Estimador Mensual de Actividad Económica (EMAE) confeccionado por el INDEC.

2. Bloque/semana: período o unidad de observación de siete u ocho días, definida según la duración del mes y que lo segmenta en cuatro bloques. De este modo se garantiza la consistencia en las comparaciones.
3. Centro Zonal: distribución geográfica interna del Banco.
4. ESA-PulsoPBA: nuevo indicador sintético que mide la evolución semanal de la actividad económica en la Provincia de Buenos Aires. La suma de las variaciones semanales de un mes es igual a la variación mensual de la serie sin estacionalidad del EMA-PBA.
5. Entrenar un modelo: en este caso, el objetivo del modelo será estimar cómo evoluciona la economía semana a semana. En el contexto de PulsoPBA, este entrenamiento algorítmico implica “mostrarle” al modelo ejemplos del pasado (los datos históricos) para que “aprenda” las relaciones entre variables bancarias y la actividad económica de la Provincia (medida con el EMA-PBA sin estacionalidad).
6. Generalización: es la capacidad que muestra un modelo de aplicar lo aprendido (en base al pasado) a estimar situaciones nuevas. Un modelo útil no se limita a repetir lo conocido, sino que puede reconocer patrones y proyectarlos en datos distintos a los utilizados para su entrenamiento. Una baja capacidad de generalización se puede deber a un sobreajuste (overfitting) o la falta de un ajuste adecuado (underfitting).
7. Predicción: una vez entrenado, un modelo puede recibir datos nuevos (por ejemplo, las transacciones de la última semana) y a partir de lo aprendido en el pasado, dar una estimación de lo que ocurre hoy o de lo que probablemente ocurrirá mañana. En el contexto de PulsoPBA:
 - a. Nowcasting: estimación en tiempo real de la actividad semanal del pasado reciente, basada en los datos bancarios más recientes.
 - b. Forecasting: proyección del futuro inmediato de la actividad económica para el bloque-semana siguiente.
8. Covariables: variables dependientes.
9. Evaluación del modelo: para asegurarse de que el algoritmo funcione bien, se compara lo que ofrece como predicción contra los valores reales observados. Si el error es bajo, el modelo se considera confiable. Las métricas de error seleccionadas para evaluar la precisión de los modelos son:
 - MAPE (mide el error porcentual promedio).
 - MAE (mide la diferencia absoluta promedio entre lo estimado y lo real).
 - RMSE (mide el error cuadrático promedio, dando más peso a errores grandes).

Fuentes de información

Se trabaja con cuatro grandes bases de datos: Movimientos, Préstamos, Comex y Consumo, cada una con características propias (resumidas en la Tabla 1):

1. Movimientos: disponible desde 2013, con frecuencia diaria y clasificación por actividad económica a dos dígitos del CLANAE, centro zonal, sector institucional. Aporta un total de 772 variables.
2. Préstamos: también disponible desde 2013, con frecuencia diaria y clasificación por actividad económica a dos dígitos del CLANAE, centro zonal, sector institucional. Aporta 94 variables.



PulsoPBA

El termómetro de actividad de la Provincia de Buenos Aires



3. Comex: disponible desde 2019, con frecuencia diaria y clasificación por actividad económica a dos dígitos del CLANAE, centro zonal, sector institucional y aporta 327 variables.
4. Consumo: disponible desde 2015, con frecuencia diaria y 21 rubros específicos (indumentaria, alimentos, etc.). No registra centro zonal, pero sí incluye variables relacionadas con clientes y compras. Tiene 57 variables y solo admite valores positivos (no negativos).

Tabla 1: características de las bases de datos insumo del algoritmo

Característica	Movimientos	Préstamos	Comex	Consumo
Registra actividad fines de semana/feriados	✗	✗	✗	✓
Fecha imputada al primer día hábil	✓	✓	✓	✗
Rubros clasificados	Actividad a 2 dígitos	Actividad a 2 dígitos	Actividad a 2 dígitos	21 rubros (indumentaria, alimentos, etc.)
Registra centro zonal	✓	✓	✓	✗
Fecha de inicio de datos	2013	2013	2019	2015
Nivel de desagregación temporal	Diario	Diario	Diario	Diario
Variables relacionadas a clientes y compras (Q_CLIENTES, Q_TRX)	✗	✗	✗	✓
Cantidad de variables	772	94	327	57
Valores nominales	✓	✓	✓	✓
Valores negativos (débitos)	✓	✓	✓	✗
Fuente de datos	Control de Gestión	Control de Gestión	Control de Gestión	Inteligencia Comercial

Fuente: elaboración propia

Todas las variables están expresadas en valores nominales. Los valores positivos son créditos y los negativos débitos. Las operaciones realizadas en monedas diferentes al Peso argentino son convertidas al tipo de cambio oficial del día de la transacción.

Semanalización de la información bancaria y del EMA-PBA

Dado que PulsoPBA se propone informar una dinámica semanal de la actividad económica, es necesario que tanto las variables bancarias (variables explicativas) como la actividad económica (variable explicada) estén expresados en la misma frecuencia semanal.

La agregación de las variables diarias en semanas presenta un desafío: los meses no empiezan el mismo día de la semana ni tienen la misma cantidad de días. Para solventar este problema metodológico, se conforma una semanalización de los meses que consiste en 4 unidades “semana-bloque” por mes, de duración variable según la cantidad de días del mes. Cada variable bancaria diaria es agregada en un bloque semanal según el esquema descrito en la Tabla 2:

Tabla 2: Método de semanalización dependiendo del mes

Días de un mes	Bloque 1	Bloque 2	Bloque 3	Bloque 4
28	7	7	7	7
29	8	7	7	7
30	8	8	7	7
31	8	8	8	7

Fuente: elaboración propia

Por otra parte, debe aumentarse la granularidad del dato de actividad económica, desde una granularidad mensual (EMA-PBA) hacia una granularidad semanal (ESA-PBA). Para ello, se adoptó un enfoque supervisado que aprovecha la riqueza de los datos bancarios. La metodología implementa el método de desagregación de Chow-Lin (1971), que garantiza que la suma de las estimaciones semanales coincide con el dato mensual publicado.

Modelos utilizados

Para realizar las predicciones se seleccionaron modelos que representan distintas tradiciones en el análisis de series de tiempo:

1. Modelos de regresiones lineales:
Se utilizó una regresión lineal similar a ARIMAX como modelo de referencia a modo de línea base. Estos modelos son simples y transparentes, y en muchos estudios han demostrado un desempeño competitivo frente a enfoques más complejos.
2. Modelos clásicos de machine learning:
Se incorporó XGBoost, un algoritmo basado en árboles de decisión que combina múltiples modelos simples para lograr predicciones más precisas. Aunque es muy potente, requiere un manejo cuidadoso para evitar el sobreajuste cuando se trabaja con grandes volúmenes de datos.
3. Modelos de Deep Learning (redes neuronales):
Se incluyeron dos arquitecturas modernas, TiDE y TSMixer, especialmente diseñadas para manejar una gran cantidad de variables. Estas técnicas basadas en redes neuronales profundas no solo aprenden a predecir, sino también a simplificar la información disponible, identificando automáticamente los patrones más relevantes.

PulsoPBA

El termómetro de actividad de la Provincia de Buenos Aires



Etapas del entrenamiento

El entrenamiento de PulsoPBA es un ciclo que se repite todos los meses y cuenta de las siguientes etapas:

1. Partición de los datos

Se realiza la partición de los datos en tres subconjuntos: entrenamiento, validación y testeo. Esta división garantiza que la evaluación del modelo se realice sobre datos no utilizados previamente durante el entrenamiento, evitando sesgos. En particular, el conjunto de validación proporciona una evaluación intermedia del modelo mientras se refinan los hiperparámetros durante la validación cruzada. Por su parte, el conjunto de testeo permanece completamente aislado hasta el final, para obtener una medida objetiva del rendimiento del modelo ya ajustado.

2. Optimización de hiper-parámetros de los modelos

Los hiper-parámetros son variables de configuración que definen la estructura y el proceso de aprendizaje de un modelo. Ejemplos habituales son la tasa de aprendizaje, el número de neuronas en una red o la profundidad de los árboles en un XGBoost. A diferencia de los parámetros internos, que se ajustan automáticamente a partir de los datos, los hiper-parámetros se fijan de antemano y determinan cómo aprende el modelo entre un conjunto de opciones posibles. Una selección adecuada de hiper-parámetros puede mejorar la precisión y generalización; mientras que valores inadecuados pueden provocar subajuste (underfitting) o sobreajuste (overfitting).

La validación cruzada es un procedimiento para estimar el rendimiento real de un modelo. Consiste en entrenar y evaluar repetidamente el algoritmo sobre distintos subconjuntos de los datos, garantizando un uso eficiente y robusto de la información disponible. En problemas de series temporales, como PulsoPBA, se emplea la modalidad de ventana móvil (rolling window). En cada iteración, el modelo se entrena con un bloque de datos históricos consecutivos y se valida en el período inmediatamente posterior. Luego la ventana de entrenamiento se desplaza hacia adelante y el proceso se repite. Este esquema preserva la secuencia temporal y permite evaluar la capacidad predictiva del modelo en contextos cambiantes, simulando el escenario de actualización continua.

El criterio principal para seleccionar los mejores hiper-parámetros fue minimizar el MAPE (Mean Absolute Percentage Error), una métrica expresada en porcentaje que mide el error relativo y evita sesgos de escala (ecuación 1), algo especialmente importante cuando se comparan períodos con distinta magnitud en la actividad económica semanal.

$$MAPE_{median} = median \left\{ \frac{1}{S} \sum_{s=1}^S \left| \frac{PulsoPBA_{w,s} - ESA_{w,s}}{ESA_{w,s}} \right|, w = 1, \dots, W \right\} \quad (1)$$

Siendo,

S: el número de semanas de la predicción

W: el número de ventanas móviles

3. Los mejores K modelos de CV evaluados en test

Tras el entrenamiento y la validación cruzada se generan múltiples modelos candidatos que demuestran ofrecer una buena calidad. En PulsoPBA, se seleccionan los K modelos con menor MAPE mediano, lo que permite comparar varios candidatos sólidos y reducir riesgos de azar. Cada modelo elegido se reentrena con todo el conjunto de entrenamiento y validación, aprovechando la máxima información disponible.

Finalmente, los modelos se evalúan en el conjunto de prueba, nunca usado en etapas previas. Allí se calculan las métricas de error, obteniendo una estimación imparcial de la capacidad predictiva real. Al igual que en la fase anterior, el error se calcula comparando las estimaciones semanales generadas por el modelo con la serie del ESA-PBA (ecuación 2).

$$MAPE_{test} = \frac{1}{S} \sum_{s=1}^S \left| \frac{\widehat{PulsoPBA}_{w,s} - ESA_{w,s}}{ESA_{w,s}} \right| \quad (2)$$

4. Competencia entre modelos y selección de los modelos para la predicción

La elección del mejor modelo se realiza mediante un proceso de filtros sucesivos que aseguran predicciones confiables y útiles. Esto son:

- Capacidad de generalización: se descartan los modelos que muestran sobreajuste, es decir, aquellos que funcionan bien en entrenamiento, pero no así en datos nuevos (conjunto de prueba).
- Cobertura: se eliminan los modelos cuyos intervalos de confianza de predicción no logran abarcar suficientemente los datos reales.
- Dirección de los cambios: se priorizan los modelos que aciertan en predecir si la economía crecerá o caerá semana a semana.

Finalmente, los modelos que superan estos filtros se ordenan según una métrica que pondera precisión, calibración y cobertura (WIS, Weighted Interval Score, ecuación 3). El mejor modelo es el que se selecciona para ser entrenado y utilizado como producto final, es decir, el que se utiliza para las predicciones de PulsoPBA.

$$WIS = \frac{1}{K+1} \{ |y - \hat{y}| + \sum_{k=1}^K \alpha_k (u_k - l_k) + \frac{2}{\alpha_k} (l_k - y) + \frac{2}{\alpha_k} (l_k - y) I_{sup} \} \quad (3)$$

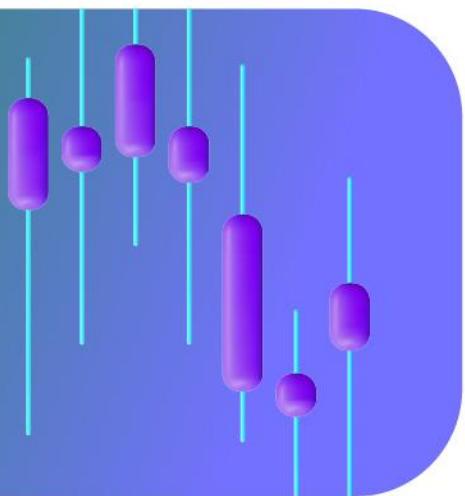
Siendo,

y: el valor observado

α_k : k-ésimo intervalo de confianza

PulsoPBA

El termómetro de actividad de la Provincia de Buenos Aires



l_k y u_k : los límites inferior y superior de α_k

\hat{y} : la mediana pronosticada

I_{inf} es la función indicador sobre el cuantil inferior $1(y < l_k)$

I_{sup} es la función indicador sobre el cuantil superior $1(y > l_k)$

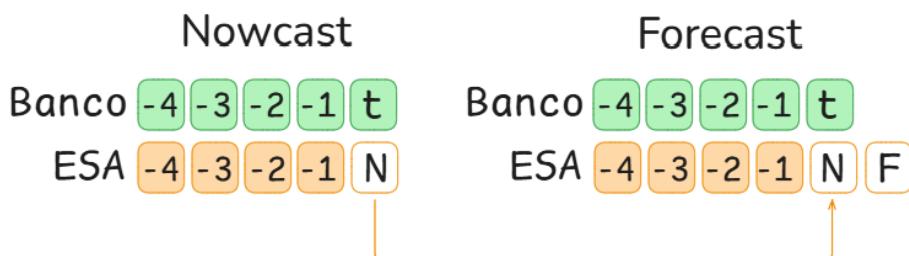
Al dividir por $K+1$, la métrica queda promediada sobre distintos intervalos y es comparable entre modelos que reporten distinto número de intervalos.

5. Reentrenamiento mensual

Cada mes, cuando se publica un nuevo dato del EMA-PBA, el modelo se actualiza con esa nueva información. Así, el sistema incorpora de manera continua las observaciones más recientes y actualiza su capacidad de predicción ajustada a la coyuntura económica.

Las predicciones

Los modelos seleccionados se ejecutan semanalmente para realizar predicciones. El modelo de nowcasting pronostica la evolución de la actividad libre de los efectos estacionales durante la cantidad de semanas que distan entre el EMA-PulsoPBA y los datos bancarios. El modelo de forecasting pronostica la actividad económica sin estacionalidad de la semana siguiente a la última pronosticada por el nowcasting.



Para más información: InvestigacionesEconomicas@bpba.com.ar

Referencias bibliográficas

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. arXiv preprint arXiv:1907.10902. <https://arxiv.org/abs/1907.10902>

Chen, Tianqi, and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–94. New York: ACM, 2016. <https://doi.org/10.1145/2939672.2939785>.

Chen, S.-A., Li, C.-L., Yoder, N., Arik, S. O., & Pfister, T. (2023). TSMixer: An all-MLP architecture for time series forecasting. arXiv preprint arXiv:2303.06053. <https://arxiv.org/abs/2303.06053>

Chow & Lin (1971) "Best Linear Unbiased Estimation of Missing Observations", Journal of the American Statistical Association, Vol. 71, No. 355.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., & Yu, R. (2023). TiDE: Time-series Dense Encoder. arXiv preprint arXiv:2304.08424. <https://arxiv.org/abs/2304.08424>

Gal, Y., & Ghahramani, Z. (2015). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (Version 6). arXiv. <https://doi.org/10.48550/ARXIV.1506.02142>