Nowcasting del PBI de Perú con Machine Learning y datos no estructurados

Juan Tenorio Wilder Perez

MEF

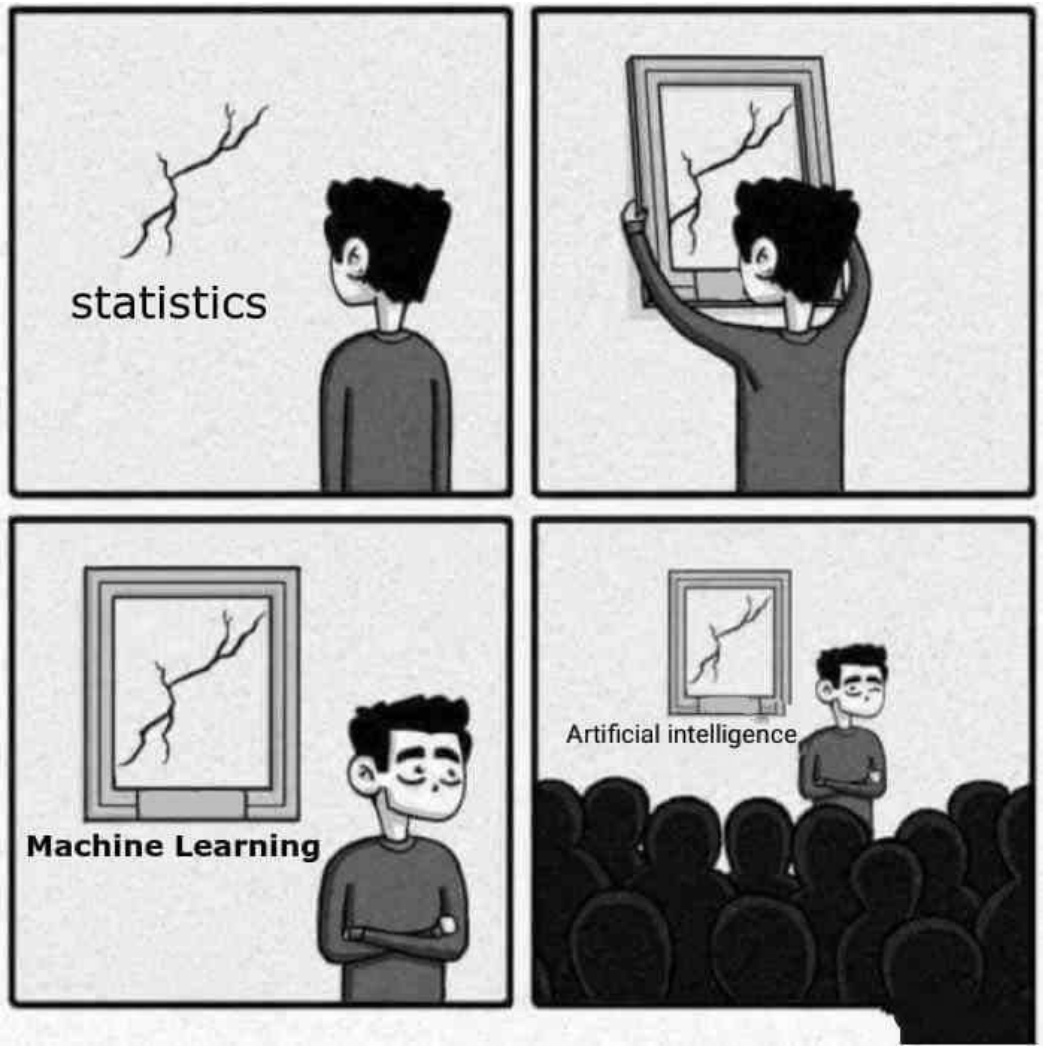
XLI Encuentro de Economistas del BCRP

*Las opiniones vertidas en este documento son de entera responsabilidad de los autores y no necesariamente reflejan las opiniones del MEF o las del BCRP

Octubre 2023

Contenido

Machine learning definition



Motivación

- Conocer la tendencia actual del estado de la economía en tiempo real, que permita tomar decisiones a los policy makers en un entorno de **alta incertidumbre** y **retraso habitual en la disponibilidad de información** de los agregados macroeconómicos.

- El continuo avance en la generación de **datos de alta frecuencia** y las **nuevas técnicas de machine learning e inteligencia artifical** han ganado gran popularidad frente al enfoque convencional de modelos tradicionales.

- En los últimos años, bancos centrales como instituciones internacionales han adoptado enfoques metodológicos que incorporan el machine learning e IA, aprovechando la **abundante cantidad de datos provenientes de motores de búsqueda y redes sociales** (Richardson y Mulder, 2018).

- Estos algoritmos suelen destacar por su **capacidad predictiva** y para **formular selecciones paramétricas en grandes conjuntos de datos**, que encuentran su base en el **entrenamiento y validación** de un porcentaje de información del modelo.

Revisión de literatura I

- **Literatura internacional nowcasting**
  - Stock y Watson (1989): primeros en proponer índices del ciclo económico utilizando modelos factoriales.
  - Giannone, Reichlin y Small (2008): evaluan el impacto marginal de la publicación de datos de frecuencia mensual en los pronósticos del PBI trimestral.
  - Evans (2005): estima en tiempo real el estado de la economía de US por medio de información en el precio de los activos mensuales.

- **Literatura ML y big data**
  - Athey (2018): explica que los algoritmos ML no supervisados significan encontrar grupos de observaciones que sean similares en términos de covarianza.
  - Athey y Markus (2017): examinan el impacto de del tipo de noticias en Google News en España en los consumidores.
  - Woloszko (2020): presenta un indicador semanal de actividad para 46 paíes de la OCDE y el G20 utilizando datos de búsqueda de Google Trends.
  - H. Varian (2014) y Niesert et al. (2020): señalan como ejemplo las consultas relacionadas a "solicitudes de desempleo" en Google Trends para pronosticar el desempleo.

Revisión de literatura II

- **Literatura nowcasting ML supervisado**
  - Richardson y Mulder (2018): indican que un modelo ML Ridge regression tiene un mejor rendimiento de predicción del PBI para Nueva Zelanda en comparación con el DFM.
  - Q. Zhang, Ni y Xu (2023): encuentran que varios algoritmos de ML superan en predicción al DFM y MIDAS para estimar el PBI de China.
  - Suphaphiphat, Wang y H. Zhang (2022): incorporando datos de Google Trends y calidad del aire en modelos DFM y ML, destacan que los ML tienen mejor desempeño en periodos de inflexión.
  - Barrios et al. (2021): desarrollan un nowcasting con ML para predecir el PBI trimestral de Belice y El Salvador, encontrando un buen poder de predicción para detectar COVID-19.

- **Literatura nowcasting Perú**
  - Kapsoli Salinas y Bencich Aguilar (2002): realiza la estimación adelantada del PBI con un modelo no lineal de redes neuronales.
  - Martınez y Quineche (2014): estiman el crecimiento del PBI en base a la producción de electricidad comparando con un modelo de RNA.
  - Pérez Forero (2018): bajo el enfoque de Varian (2014) a través del Gibbs-Sampling y un prior spike-and slab calcula la probabilidad de inclusión de los mejores indicadores para estimar el PBI .

Content

Metodología

- **Modelo benchmark autoregresivo** un modelo AR para el crecimiento mensual del PBI (y_t), el cual refleja el valor de una variable en función de sus propios valores previos. Un modelo de orden 1, presenta la siguiente estructura:

$$y_t = \beta_0 + \beta_1 y_{t-1} + e_t \tag{1}$$

donde e_t captura la aleatoriedad del modelo.

- **Modelo de factores dinámicos:** se estima un modelo DFM canónico que siguiendo a Evans (2005), se puede describir de la siguiente manera:

$$x_t = C_0 f_t + e_t \qquad e_t \sim N(0, R) \tag{2}$$

$$f_t = \sum_{j=1}^{p} A_j f_{t-j} + u_t \qquad u_t \sim N(0, Q_0) \tag{3}$$

donde la ecuación 2 es de medida y la ecuación 3 de transición, permitiendo al factor no observable f_t comportarse como un VAR.

Content

LASSO regression

El modelo LASSO, introducido por Tibshirani (1996), utiliza una penalización basada en la suma de los valores absolutos de los coeficientes de las variables predictoras. **Esta penalización tiene la propiedad de forzar algunos coeficientes a alcanzar exactamente el valor cero**.

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{4}$$

el cambio radica en el hiperparámetro λ lo que permite simplificar el modelo, resultando una selección automática de un **subconjunto de variables predictoras más relevantes y en la eliminación de variables menos significativas**.

Ridge regression

El modelo Ridge, se define por agregar una penalización basada en la suma de los cuadrados de los coeficientes de las variables predictoras. **Esta penalización fuerza los coeficientes a ser muy pequeños, evitando que tomen valores extremadamente altos** y, por lo tanto, reduciendo la influencia de variables menos relevantes.

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{5}$$

λ es el hiperparámetro de penalización que controla la magnitud de la regularización. La suma de los términos β_j^2 en la penalización impide que los coeficientes alcancen valores grandes, **contribuyendo así a la estabilidad y reducción del riesgo de sobreajuste (overfitting)**.

Elastic Net Regression

El modelo Elastic Net, combina de forma apropiada las limitaciones del modelo LASSO y Ridge. En particular Zou y Hastie (2005), mencionan que su ventaja radica en **corregir el modelo cuando el número de regresores es mayor al de observaciones ($p > n$)** lo que mejora la agrupación de variables.

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)\beta_j^2 \right) \right) \tag{6}$$

donde λ es el hiperparámetro de penalización global y α es el hiperparámetro que controla la mezcla entre las penalizaciones de LASSO y Ridge. La combinación de ambas penalizaciones **permite un mayor grado de flexibilidad al seleccionar variables y alinear los coeficientes**.

Adaptive Lasso Regression

Siguiendo a Zou (2006), el modelo Adaptive LASSO, se presenta como una variante del modelo LASSO, que introduce un enfoque de regularización que **ajusta de manera adaptativa la magnitud de las penalizaciones para cada coeficiente de las variables predictoras**. Esta adaptación permite que las penalizaciones sean diferentes, lo que puede resultar en una selección más precisa de variables relevantes

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right) \tag{7}$$

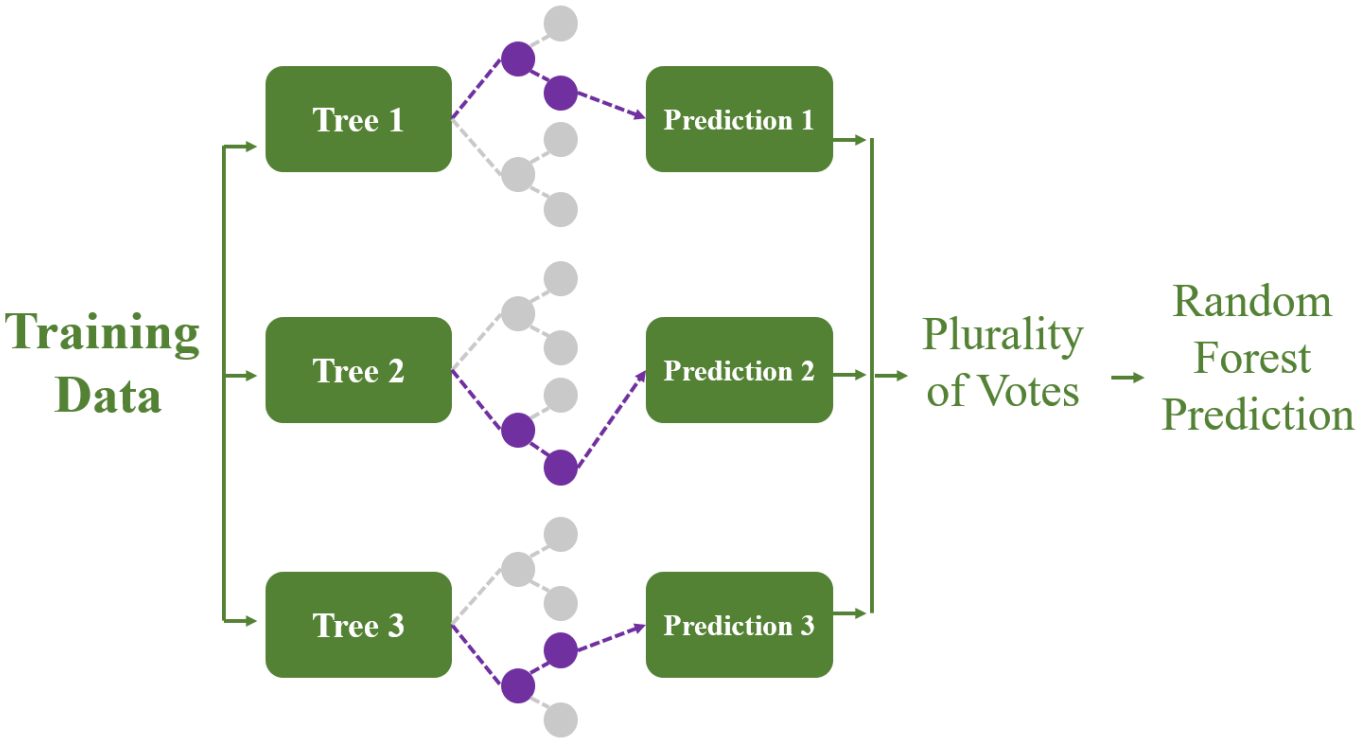
donde λ es el hiperparámetro de penalización, w_j es el factor de adaptación para el coeficiente β_j y depende de la implementación específica en el algoritmo, **en este caso se usan los pesos del Ridge regression**.

Content

Random forest

Los árboles de decisión son algoritmos ML que **representan decisiones y acciones** en forma de un árbol con nodos internos (train-set) y ramas (predicción).

Figure: Simple Representation of the Random Forest Algorithm

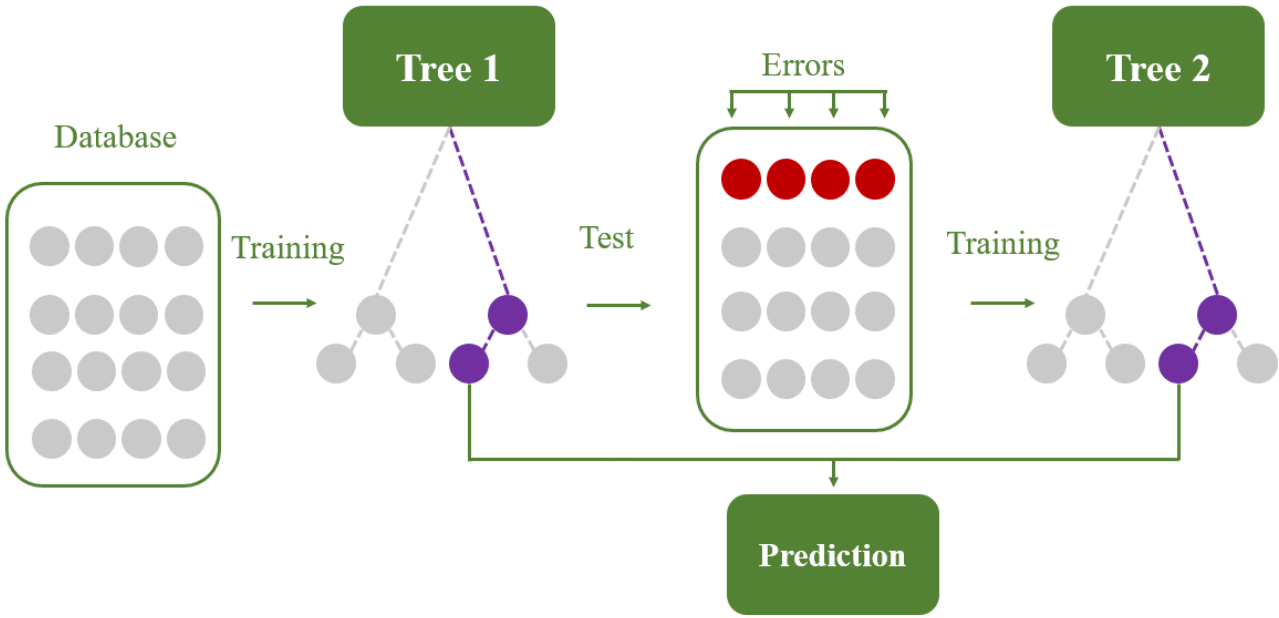


Source: Own elaboration

Gradient Boosting Machine

Estos modelos **entrenan árboles utilizando los errores del conjunto acumulado de predicciones débiles de manera que proporcione una mejora progresiva** en el rendimiento de predicción (Natekin y Knoll, 2013).

Figure: Simple Representation of the Gradient Boosting Machine Algorithm



Source: According to Boehmke y Greenwell (2020)

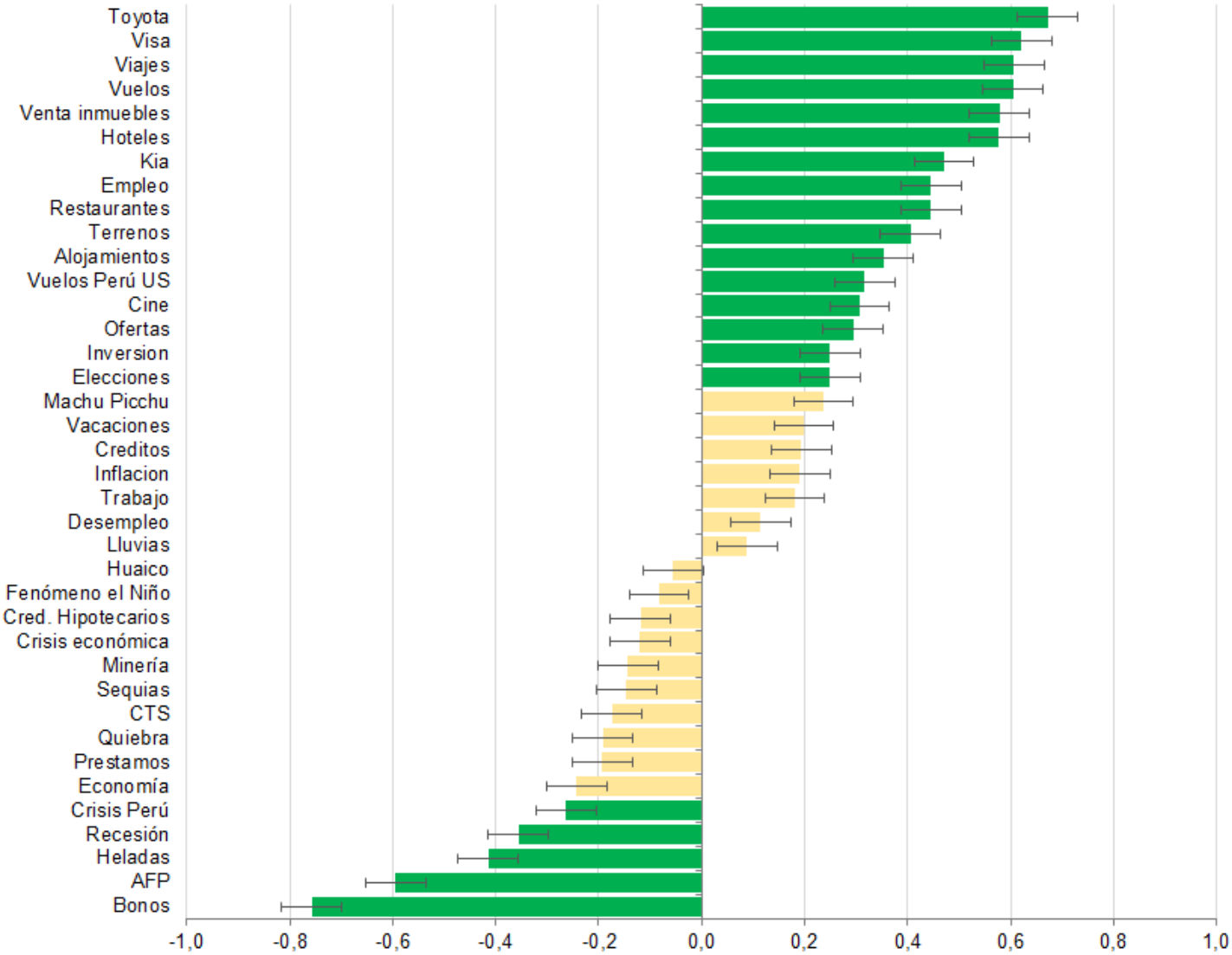
Datos no estructurados

Table: Example list of no structured variables included in the model

Unstructured variable details		
Units of Measure	**Frequency**	**Source**
Search Index (0 to 100)	Daily	Google Trends
Variables		
1.- Searched Words on Economic		
Inflación	Recesión	
2.- Searched Words on Consumption		
kia	toyota	Cinema
Restaurantes	Créditos	Préstamos
Hipotecarios	Ofertas	
3.- Searched Words on Labor Market		
Empleo	Desempleo	Trabajo
4.- Searched Words on Sectorial Industry		
Minería	Inversión	

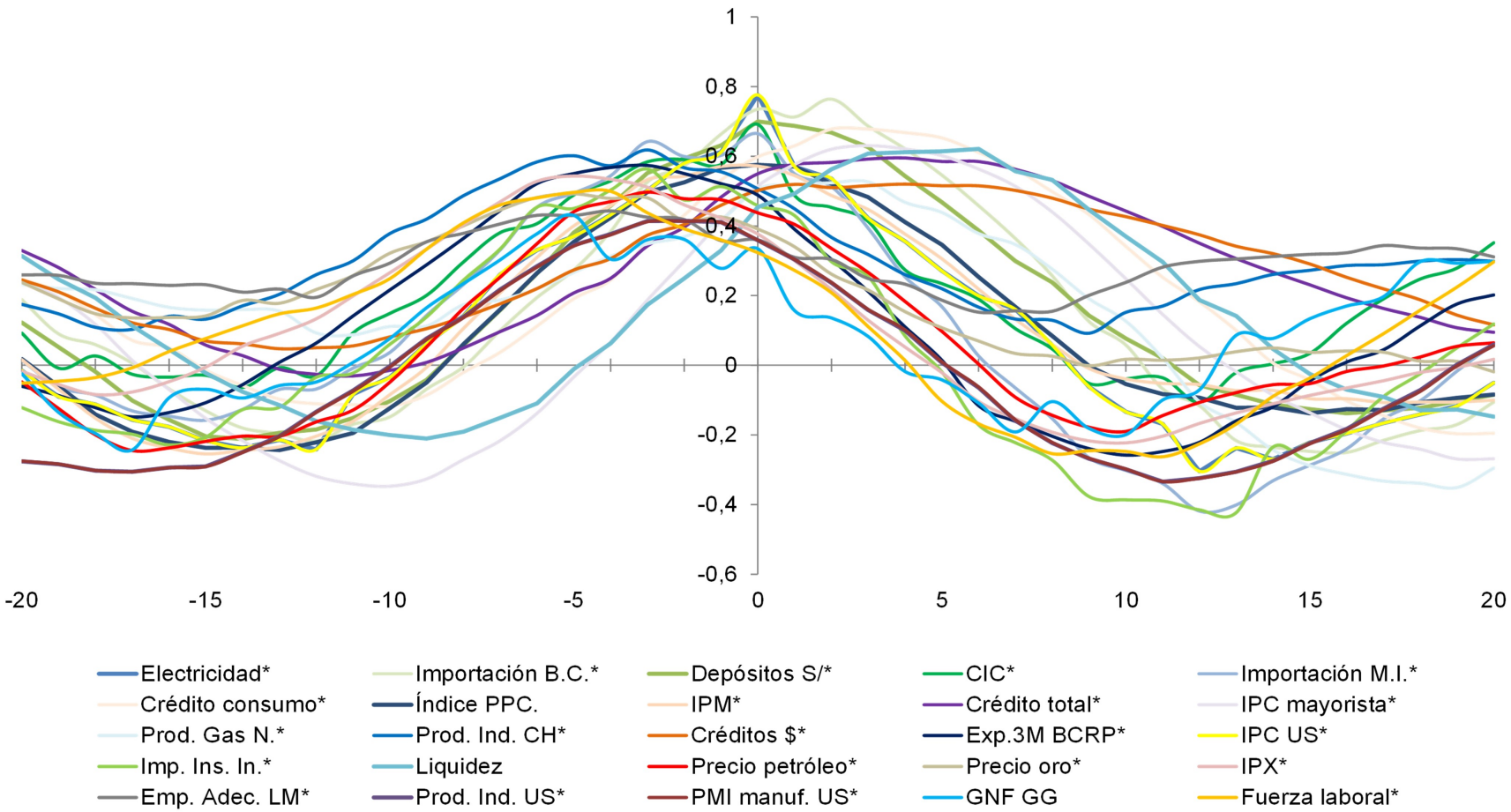
Source: Own elaboration

Figure: Correlations of the main nonstructured variables



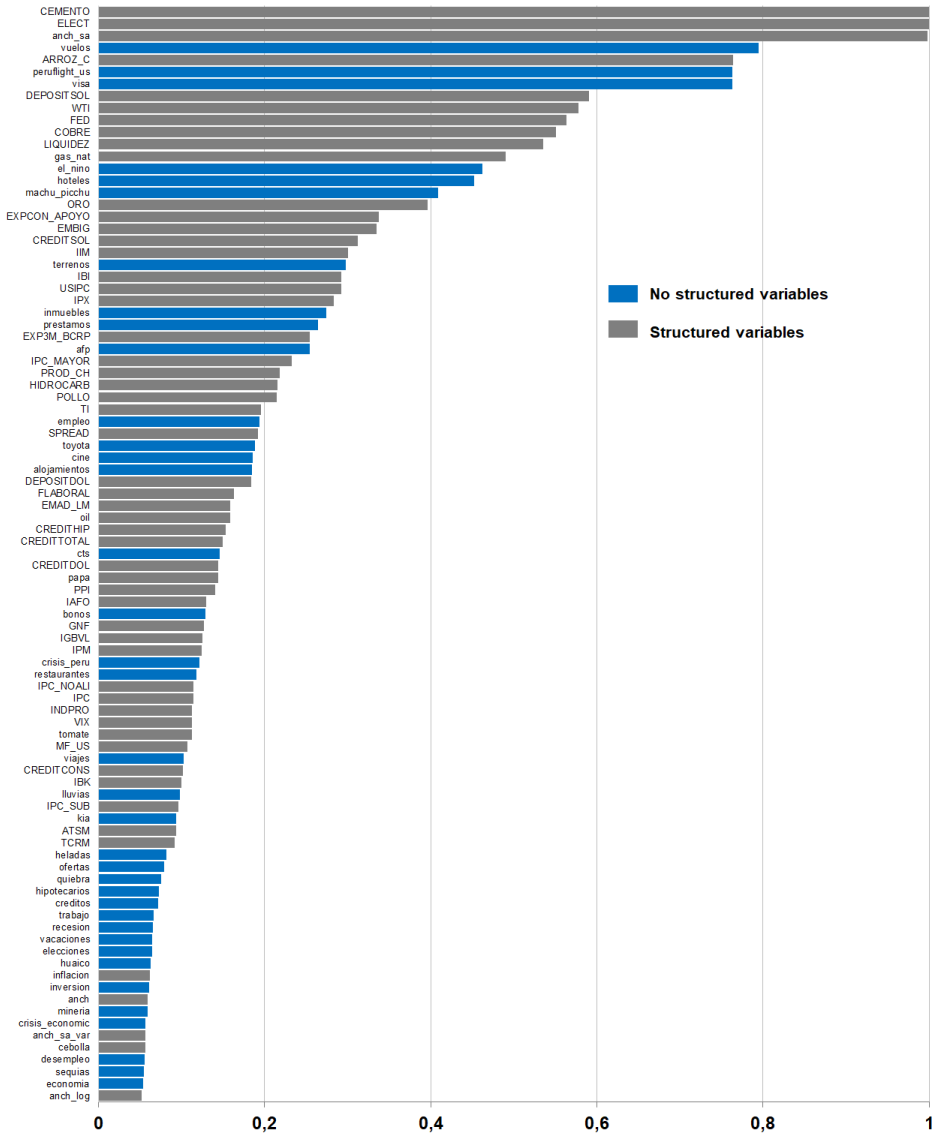
Datos estructurados

Figure: Dynamic correlations of the main variables



Source: Own elaboration y (*) simboliza las variables causales a la Granger

Figure: Gibb sampling (2004-2023) - probability of inclusion in optimal model



Evaluación de previsión

- Disponemos finalmente de un conjunto total de **91 predictores que abarcan desde enero de 2008 hasta mayo de 2023**. La evaluación y selección de los predictores óptimos se llevará a cabo de manera independiente para cada algoritmo ML.
- **Estrategia de evaluación de proyecciones:** El método que evaluará la precisión en la proyección de cada modelo se realizará por medio del error cuadrático medio (RMSE), siguiendo la ecuación:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{8}$$

- Seguido de esta evaluación emplearemos el método de Diebold y Mariano (1995), para determinar si las proyecciones generadas por cada modelo de ML difieren significativamente en comparación con el modelo *benchmark*.

Estimación del modelo y calibración de hiperparámetros

- Se divide el conjunto de datos en **tres partes clave: entrenamiento, validación y prueba**. Inicialmente, el modelo se entrena utilizando los datos del conjunto de entrenamiento (*in-sample*), los cuales minimicen el error cuadrático medio.
- Por validación cruzada (CV) una vez identificados los valores óptimos, se **emplea el conjunto de prueba para evaluar la capacidad predictiva del modelo fuera del conjunto de entrenamiento** (*out-sample*)
- CV consiste **en entrenar y validar el modelo en cinco ventanas (5 folds)**, utilizando cada fragmento de datos como conjunto de validación y los restantes como conjunto de entrenamiento en cada iteración.

Table: Strategy of testing estimations

	Training dataset 2008m1-2014m08				**Testing set** 2014m09-2023m5
	⟷				↔
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	

Source: Own elaboration

- Una estrategia en la configuración de los modelos para **prevenir el overfitting y riesgo de sobre-entrenamiento**, implica la limitación de los hiperparámetros dentro de los rangos recomendados por la literatura (Zou, 2005).

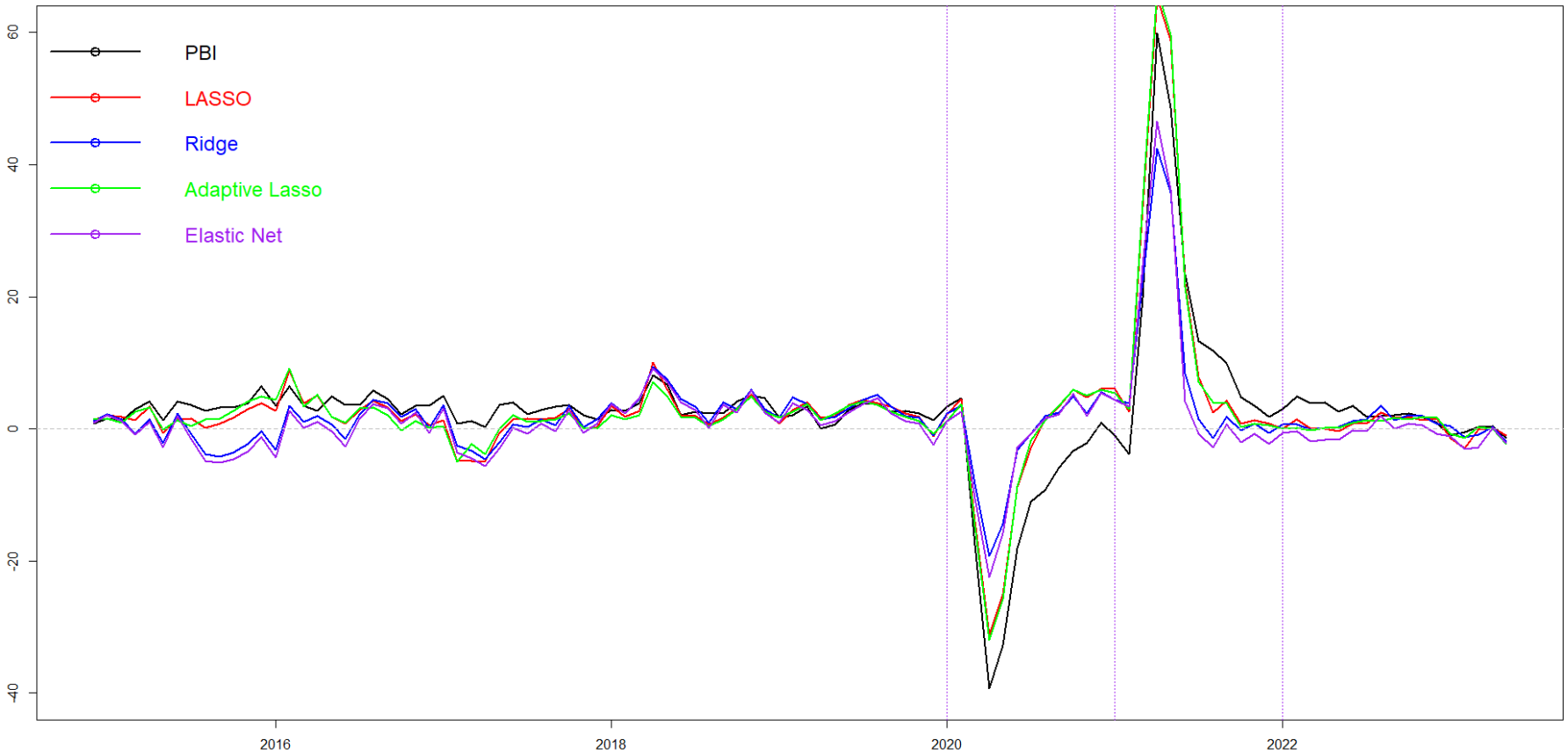
Table: Priors and hyperparameter ranges

Model	Hyperparameter	Range	Optimised Value
Lasso	Lambda	0.001 to 0.009	0.007
Ridge	Lambda	0.01 to 0.09	0.310
Elastic Net	Alpha	0.1 to 0.9	0.500
	Lambda	0.01 to 0.09	0.040
Adaptive Lasso	Lambda	0.01 to 0.09	0.670
	Omega	0.1 to 0.9	0.340
Random Forest	#árboles	1 to 400	281
Gradient Boosting Machine	# árboles	1 to 5000	19
	Distribución	Normal	Bernoulli
	Shrinkage	0.001 to 0.009	0.300

Source: Own elaboration

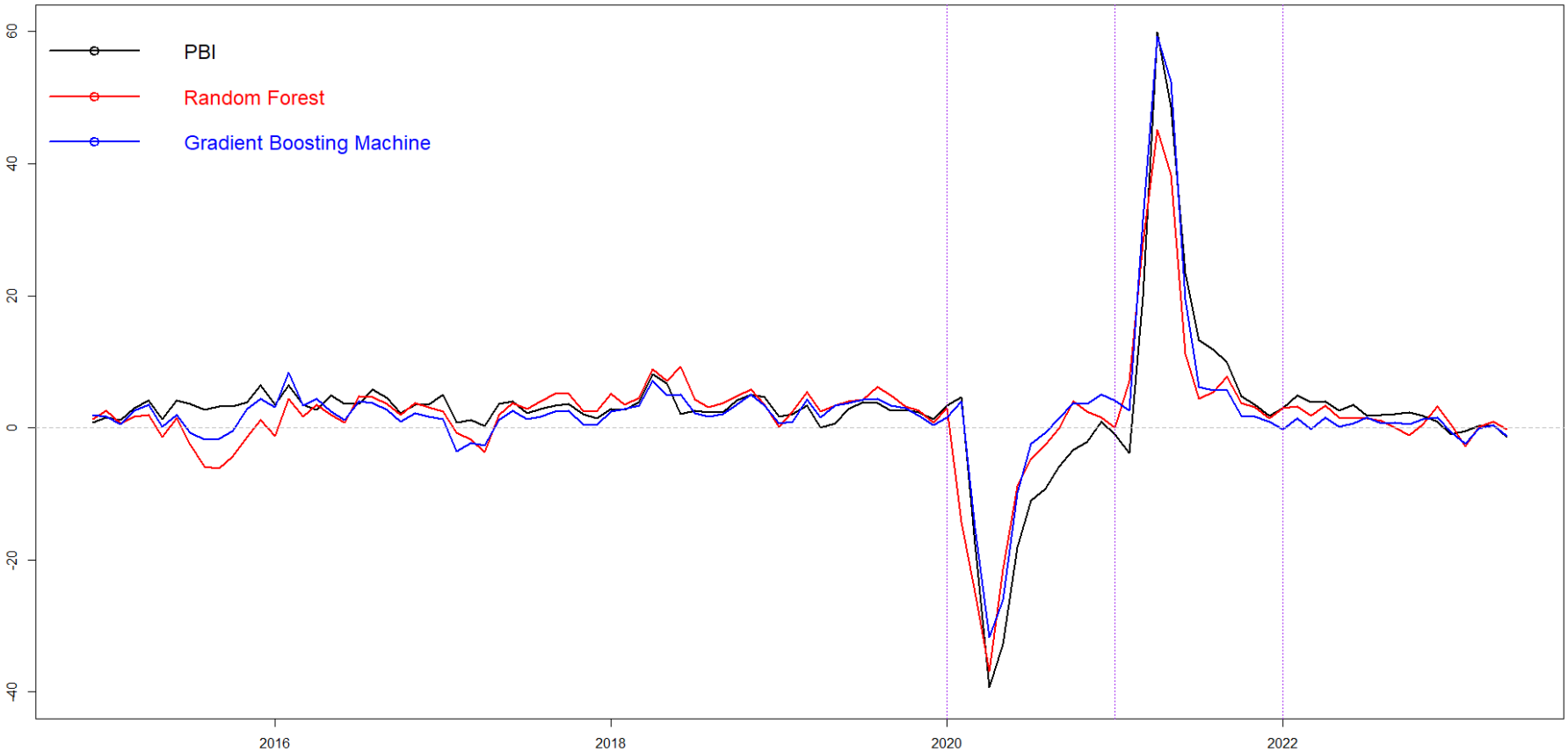
Resultados PBI vs nowcasting Machine Learning I

Figure: ML model projection and GDP



Resultados PBI vs nowcasting Machine Learning II

Figure: Tree models projection and GDP



Comparación de modelos

- En términos de RMSE, los modelos ML **logran ser los más pequeños** en comparación con el modelo *benchmark* y el DFM. Destacando el GBM, LASSO y Enet.
- Además, el estadístico DM confirma que la mayoría de modelos ML son estadísticamente **significativos**, respaldando la eficacia de estas metodologías, en línea con investigaciones previas.

Table: Evaluation of model and benchmark forecasts
2014m09-2023m12

Model	RMSE	RMSE (Rel. to AR)	p-value
Lasso	0.26	0.79	0.014
Ridge	0.34	0.68	0.043
Elastic Net	0.28	0.55	0.039
Adaptive Lasso	0.68	0.71	0.126
Random Forest	0.45	0.76	0.089
Gradient Boosting Machine	0.17	0.21	0.016
Dynamic Factor Model	1.39	1.01	0.105
AR	2.55	0.00	

Source: Own elaboration

Consistencia y robustez

Para evaluar el modelo y determinar si las proyecciones de los modelos de ML **aportan algún valor adicional a los pronósticos del indicador mensual en comparación con un segundo modelo benchmark**, se emplea un enfoque que incluye la estimación de un DFM del PBI con producción de electricidad, siguiendo de cerca a Romer y Romer (2008).

$$y_t = \beta_1 DFME_t + \beta_2 ML_{it} + e_t \tag{9}$$

Table: β_2^e value and validation criteria

Models	Estimated value	AIC	p-value	p-value (DM)
Lasso	0.714	520.32	0.000	0.079
Ridge	0.936	554.73	0.000	0.057
Elastic Net	0.839	549.80	0.000	0.055
Adaptive Lasso	0.703	517.49	0.000	0.046
Random Forest	0.783	534.20	0.000	0.049
Gradient Boosting Machine	0.810	492.09	0.000	0.041

Source: Own elaboration

Conclusiones y agenda pendiente

- Los resultados indican que las **predicciones de los modelos de machine learning son más sólidas en comparación con el modelo benchmark**. En concreto, los modelos Random Forest, Gradient Boosting Machine y Adaptive Lasso muestran un rendimiento con una capacidad superior.

- Los resultados de evaluación y ejercicio de coherencia **muestran evidencias de la contribución positiva de los modelos ML** y los datos de sentimiento mejoran significativamente la precisión del modelo permitiendo la detección temprana de períodos de alta volatilidad, un aspecto que los modelos convencionales a menudo no logran captar.

- Nuestros resultados arrojan luz sobre la superación del machine learning respecto a los modelos AR y DFM en predicción, lo que abre una **nueva agenda de enfoques sobre mejoras de la previsión de otras variables macroeconómicas relevantes como el consumo, empleo, inversión, entre otras.**

Nowcasting del PBI de Perú con Machine Learning y datos no estructurados

Juan Tenorio Wilder Perez

MEF

XLI Encuentro de Economistas del BCRP

*Las opiniones vertidas en este documento son de entera responsabilidad de los autores y no necesariamente reflejan las opiniones del MEF.

Octubre 2023