Optimización automatizada de preselección de variables y modelos predictivos de *machine learning* para el *forecasting* del PBI peruano

Walter Ruelas-Huanca
Banco Central de Reserva del Perú

Bruno Gonzaga
Banco Central de Reserva del Perú

Andrew Garcia
Emprendedor - IA y Software

Encuentro de Economistas
Octubre 2024

Motivación (1/3)

- Usualmente, **la proyección de variables económicas implica asumir ciertas especificaciones o supuestos** antes de entrenar y testear uno o varios modelos.

- Por ejemplo: la decisión de reducir el conjunto de variables relevantes, la elección de aplicar ciertos procedimientos de “limpieza” de datos (desestacionalización / *outliers*), el mejor modelo para predecir la variable objetivo (y cómo se define esta variable), entre otros.

Motivación (2/3)

- Supongamos que optamos por reducir la dimensionalidad de un conjunto de datos y luego entramos un modelo que prediga de mejor manera a la variable objetivo. ¿El mejor modelo será independiente de la técnica de reducción de dimensionalidad que se ha aplicado?

Motivación (3/3)

- Creemos que nuestra **principal contribución** a la literatura recae sobre la definición y aplicación de una arquitectura capaz de “hiperparametrizar” elecciones que son usualmente asumidas por el usuario.
- Una **contribución secundaria** está relacionada a la posibilidad de ampliar la arquitectura a otras especificaciones como: definición de la variable objetivo, elección de técnicas de desestacionalización, entre otros.

Literatura (1/1)

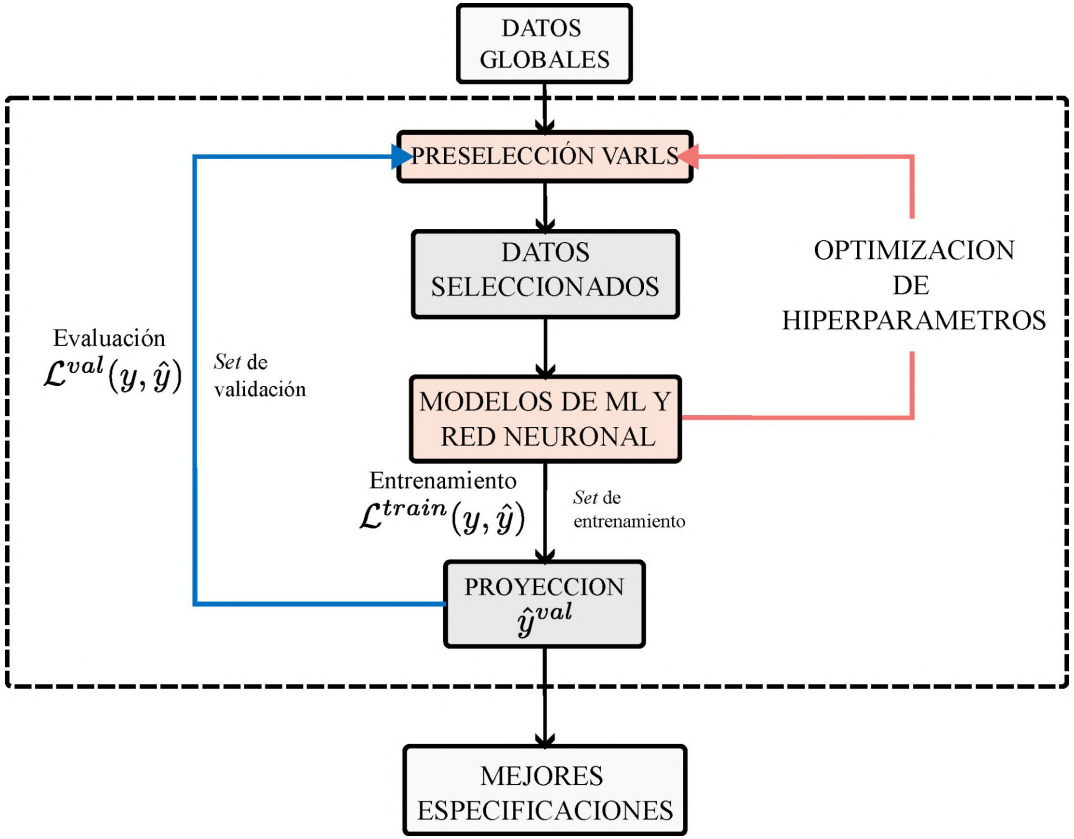
Nowcasting y *Forecasting* del PBI:

- Con modelos de ML: Tenorio y Pérez (2024) en Perú, Silva et al. (2024) en Brasil, Yang et al. (2024) en China.
- Con redes neuronales: Cai et al. (2021), Jung et al. (2018).
- Con métodos de ensamblaje: LIN Jian (2005), Longo et al. (2022).

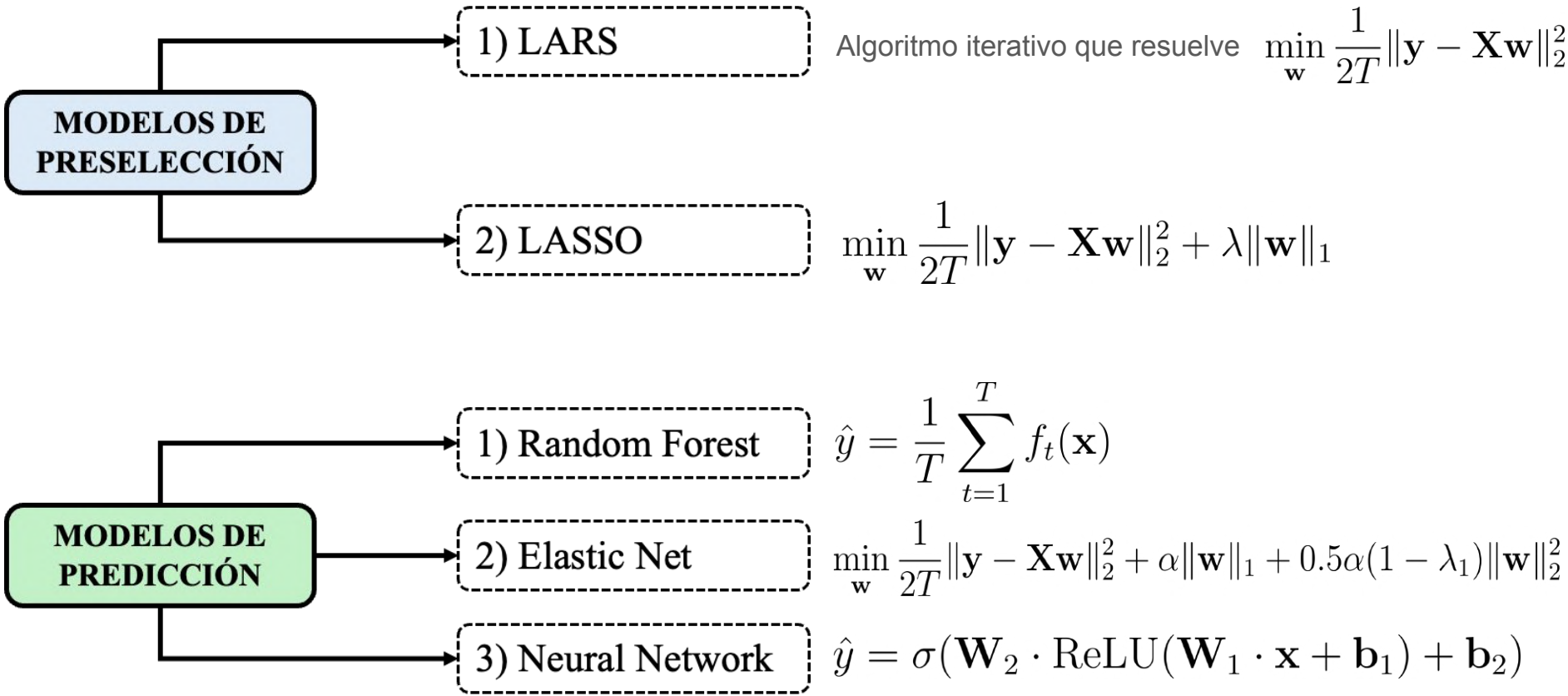
Preselección de datos es relevante:

- Li et al. (2022a), Hmamouche et al. (2017), Garcia y Santana (2024), Ben Ishak (2016).

Arquitectura (1/2)



Arquitectura (2/2)



Metodología (1/5)

- **Tree-structured Parzen Estimator (TPE)**: Técnica de optimización bayesiana, la cual usa estimadores de Parzen (kernel) para modelar creencias acerca de los hiperparámetros y guiar su búsqueda mediante una función de adquisición, la cual pondera exploración y explotación.

- Dicha función de adquisición está basada en el ratio de dos distribuciones de probabilidad: una de configuraciones de hiperparámetros que causan bajas pérdidas (buenas configuraciones) y otra en configuraciones de hiperparámetros que causan altas pérdidas (malas configuraciones).

Metodología (2/5)

- Este enfoque sigue a Bergstra et al (2011) y Watanabe (2023).

- El objetivo es encontrar $\hat{x} = \arg\max_x EI_{y^*}(x)$

  donde EI o *Expected Improvement*: $EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(y|x)\, dy$

  - x son los hiperparámetros a optimizar.
  - y^* es el mejor valor observado de la función de pérdida hasta el momento.
  - $p(y|x)$ es la distribución de la función de pérdida dado un conjunto de hiperparámetros.

Metodología (3/5)

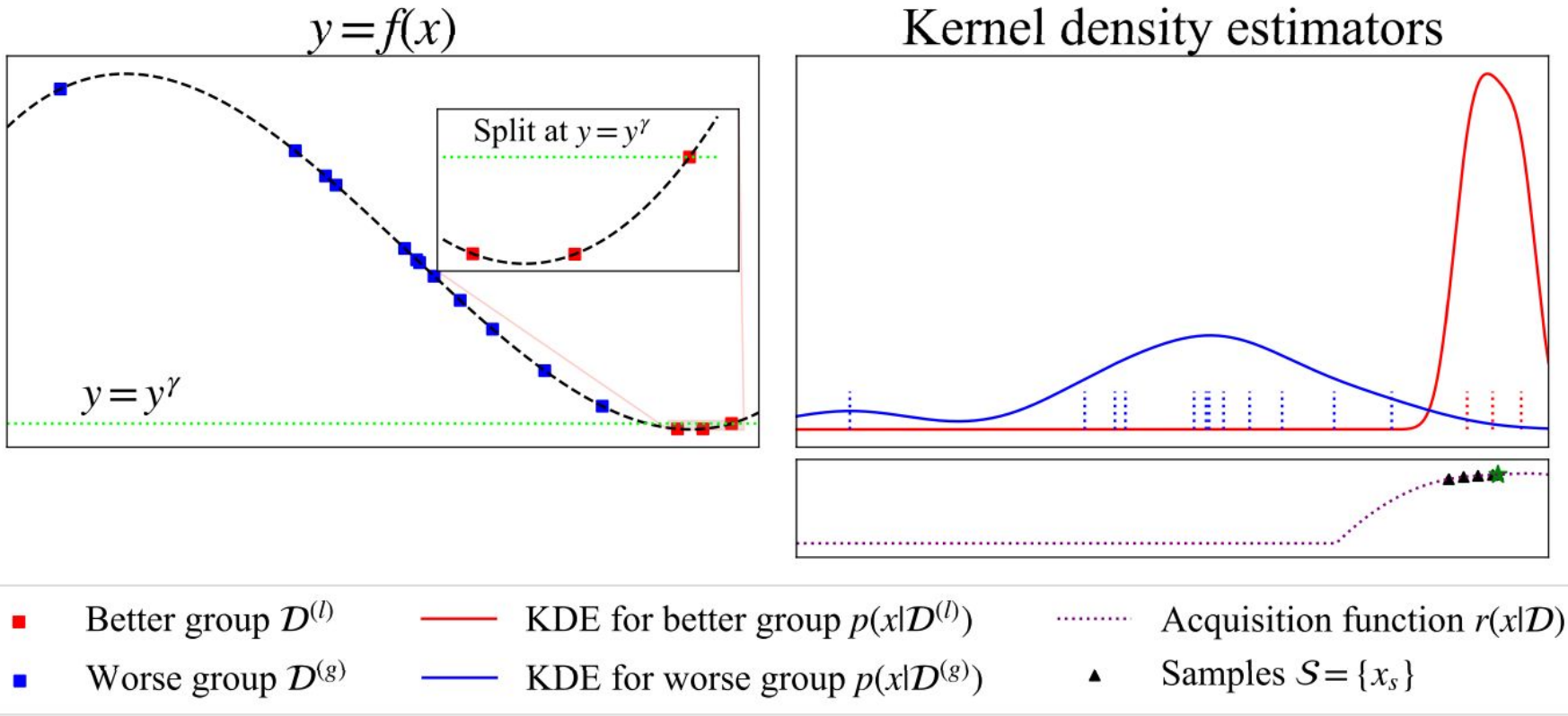
- Si se tiene que $p(x|y) = \begin{cases} l(x) & \text{si } y < y^* \\ g(x) & \text{si } y \geq y^* \end{cases}$

  Entonces:

$$EI_{y^*}(x) \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1} \quad \text{con} \quad p(y < y^*) = \gamma$$

  donde

  - $l(x)$ representa la densidad de probabilidades de los hiperparámetros que conducen a bajas pérdidas (buenas configuraciones).
  - $g(x)$ representa la densidad de probabilidades de los hiperparámetros que conducen a altas pérdidas (malas configuraciones).



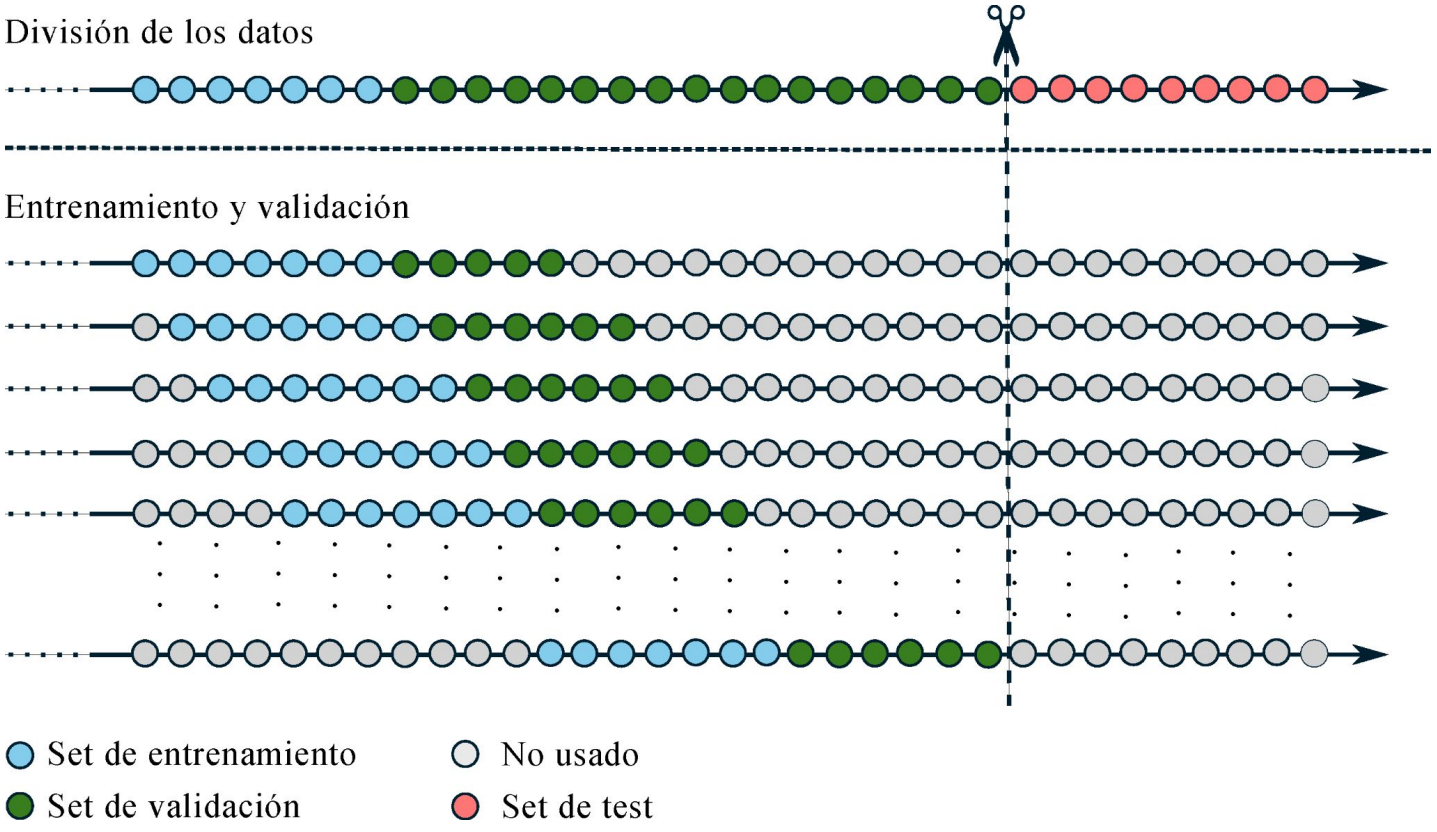
Fuente: Watanabe, S. (2023). Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*.

Metodología (4/5)

ESPACIO DE BÚSQUEDA

Etapa	Modelo	Hiperparámetro	Espacio de Búsqueda
Preselección	LASSO	Magnitud de penalización	$\log(\text{uniform}(-5, 2))$
	LARS	N° coef. no nulos	$\text{int}(\text{quniform}(10, 50, 1))$
	Ninguno	-	-
Forecasting	Random Forest	N° de árboles	$\text{int}(\text{quniform}(10, 300, 1))$
		Profundidad máx. de árboles	$\text{int}(\text{quniform}(2, 32, 1))$
		N° mín. de muestras necesarias para dividir un nodo	$\text{int}(\text{quniform}(2, 16, 1))$
		N° mín. de muestras necesarias en una hoja	$\text{int}(\text{quniform}(1, 16, 1))$
	Elastic Net	Peso total de la regularización	$\log(\text{uniform}(-5, 2))$
		Peso de la regularización L1	$\text{uniform}(0, 1)$
	Red Neuronal	Tasa de aprendizaje	$\log(\text{uniform}(-5, -2))$
		N° de épocas	$\text{int}(\text{quniform}(20, 100, 10))$

Metodología (5/5)



Datos (1/1)

BASE DE DATOS, SEGÚN GRUPOS

Grupos de variables	
PBI total	Confianza del consumidor
PBI por sectores	Indicadores de crédito
Electricidad	Indicadores bancarios
Cemento	Indicadores del LBTR
Hidrocarburos	Tasas de interés
IPC y componentes	Indicadores de renta fija y variable
Producción de alimentos	Tipos de cambio
Indicadores de recaudación fiscal	Riesgo país
Volumen de export. e import.	Búsquedas en Google Trends
Indicadores laborales	Temperatura marina
Expectativas empresariales	Indicadores económicos mundiales

Resultados (1/3)

MEJOR ESPECIFICACIÓN, SEGÚN HORIZONTE DE PROYECCIÓN

Etapa	Modelos	Hiperparámetro	$h = 1$	$h = 2$	$h = 3$
Preselección	LARS	Número de coeficientes no nulos seleccionados	19	15	17
Proyección	Random Forest	Número de árboles en el bosque	16	213	43
		Profundidad máxima de los árboles	4	5	11
		Número mínimo de muestras necesarias para dividir un nodo	10	2	3
		Número mínimo de muestras necesarias en un nodo hoja	3	2	1

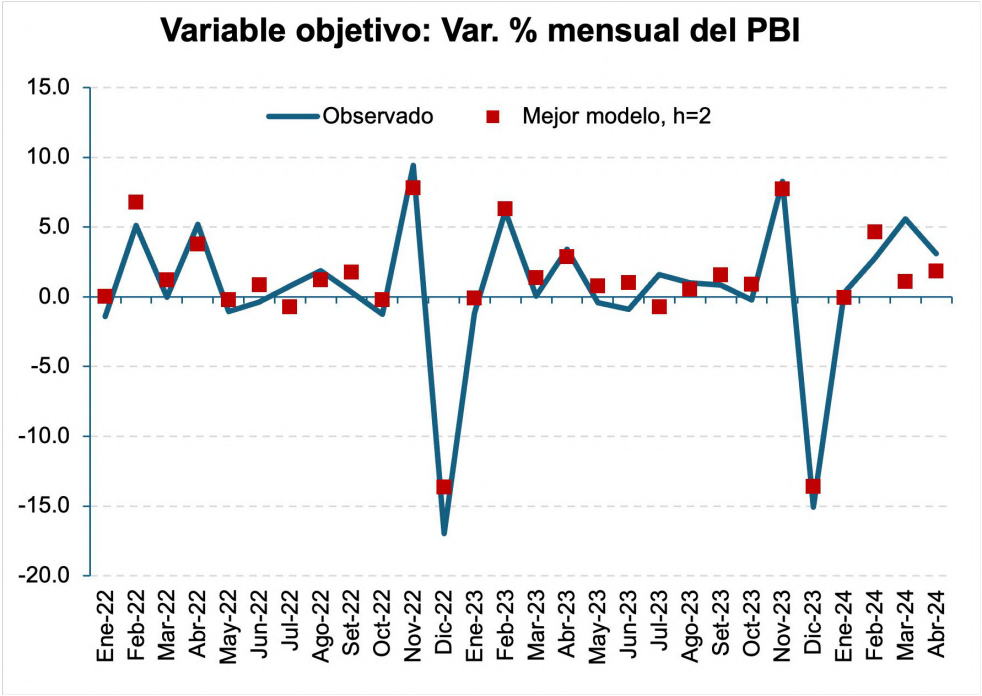
Resultados (2/3)

RAÍZ DEL ERROR CUADRÁTICO MEDIO (RECM)

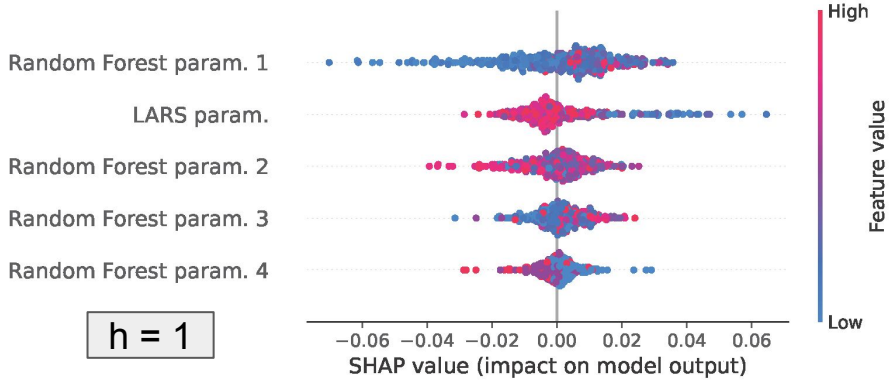
Var. objetivo: Var. % mensual del PBI

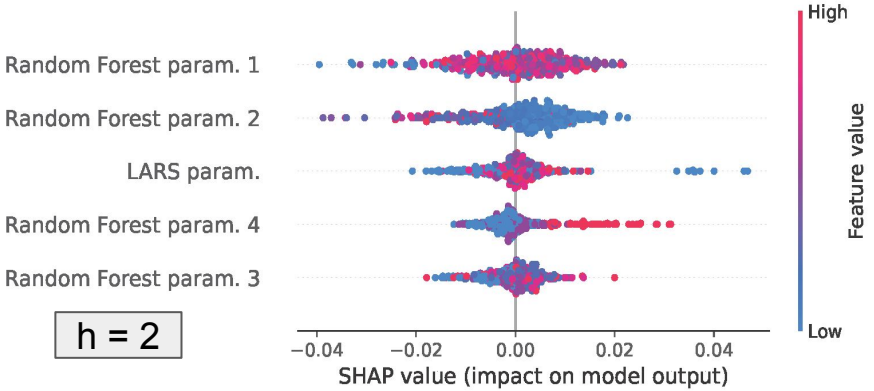
Modelo	RECM		
	$h = 1$	$h = 2$	$h = 3$
Random Forest, LARS*	1,5492	1,6218	1,5644
Random Forest, LARS**	1,6054	1,6293	1,6047
Random Forest, LARS***	1,6083	1,6294	1,6311
Random Forest, LARS****	1,6089	1,6312	1,6321
Random Forest, LARS*****	1,6204	1,6377	1,6374
AR(1)	5,7797	5,4066	5,7798
AR(2)	5,5930	5,4576	5,6414
AR(4)	5,2926	5,5616	5,8572
AR(6)	5,3453	5,7060	5,7302
AR(8)	5,2242	6,2300	5,6470
VAR(1)	5,4289	6,2080	5,5388
VAR(2)	5,6597	6,7812	6,4627
VAR(4)	5,3402	6,3744	7,1731
VAR(6)	4,7749	7,0005	8,2516
VAR(8)	5,3478	7,3361	6,8468

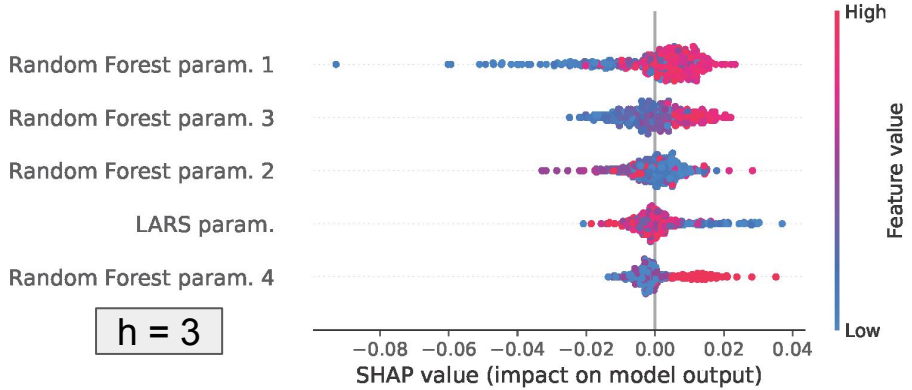
*Primera mejor especificación, ** segunda mejor especificación, etc.



Resultados (3/3)







- RF param.1: núm. de árboles
- RF param.2: profundidad máx. de árboles
- RF param.3: núm. mín. de muestras necesarias para dividir un nodo
- RF param.4: núm. mín. de muestras necesarias en un hoja
- LARS param.: núm. de coef. no nulos

Discusión (1/1)

- La preselección de variables mejora el desempeño del modelo en comparación con su ausencia.

- La mejor especificación es LARS + Random Forest, y su desempeño fue mejor al de modelos tradicionales como VAR o AR. La red neuronal no obtuvo buenos resultados posiblemente debido a su simplicidad.

- La arquitectura propuesta sirve como referencia para futuras mejoras, adiciones y/o complejizaciones.

Agenda futura (1/1)

- Afinar el *benchmark*.

- Complejizar la arquitectura incluyendo más modelos o perfeccionando los ya disponibles (RRNN), o también incorporando más etapas.

Optimización automatizada de preselección de variables y modelos predictivos de *machine learning* para el *forecasting* del PBI peruano

Walter Ruelas-Huanca
Banco Central de Reserva del Perú

Bruno Gonzaga
Banco Central de Reserva del Perú

Andrew Garcia
Emprendedor - IA y Software

Encuentro de Economistas
Octubre 2024