Modelos de Machine Learning (ML) basados en indicadores macroeconómicos para predecir la pérdida esperada en el sistema crediticio de México

Andrea Rancaño Botaya José Angel Alonso Prieto
Eduardo Daniel Morales Sánchez

Resumen

Este estudio evalúa si los modelos de aprendizaje automático (ML), entrenados con indicadores macroeconómicos, pueden anticipar de manera precisa y oportuna la pérdida esperada en distintos segmentos del sistema crediticio mexicano. Se utilizaron datos provenientes de la CNBV, Banxico, INEGI, IMSS y Google Trends, abarcando variables tradicionales y no convencionales. El análisis considera tanto modelos clásicos de series de tiempo (ARIMA, SARIMA, VAR, Prophet) como enfoques más flexibles basados en árboles de decisión (Bagging, Random Forest, XGBoost). Los resultados muestran que los modelos basados en árboles, al incorporar dinámicas temporales y variables exógenas, superan consistentemente a los modelos tradicionales en términos de precisión predictiva. Este enfoque ofrece una herramienta para el monitoreo del riesgo crediticio y la formulación de políticas macroprudenciales para las instituciones financieras.

Introducción

El monitoreo del riesgo crediticio es esencial para preservar la estabilidad del sistema financiero. En contextos de alta incertidumbre macroeconómica, la capacidad de anticipar aumentos en la pérdida esperada de los portafolios crediticios se convierte en una ventaja clave tanto para instituciones financieras como para autoridades regulatorias. Tradicionalmente, este monitoreo se ha basado en modelos econométricos lineales y en evaluaciones periódicas de desempeño. Sin embargo, la creciente disponibilidad de datos de alta frecuencia y herramientas de ML permite explorar alternativas más flexibles y oportunas.

Pregunta de investigación

Este trabajo busca responder la siguiente pregunta de investigación:

¿Pueden los modelos de Machine Learning, entrenados con indicadores macroeconómicos, anticipar de manera precisa y oportuna la pérdida esperada futura del sistema crediticio mexicano?

Hipótesis

Si las condiciones macroeconómicas tienen un efecto sistemático sobre la evolución del riesgo crediticio, entonces los modelos de aprendizaje automático, al incorporar esta información, deberían ser capaces de anticipar cambios en la pérdida esperada con alta precisión.

Motivación

Al permitir una identificación temprana de señales de deterioro crediticio, estos modelos se convierten en una herramienta clave tanto para las instituciones financieras como para las autoridades regulatorias, al facilitar la toma de decisiones informadas y el diseño de políticas macroprudenciales orientadas a preservar la estabilidad financiera.

Datos

En este análisis se utilizaron diferentes variables para entender cómo evoluciona el riesgo crediticio y qué factores pueden influir en él. Por un lado, consideramos las pérdidas esperadas en distintos tipos de créditos (todos de stage 1); por otro, incluimos variables macroeconómicas que ayudan a explicar los cambios en estos riesgos. A continuación, se presentan las variables dependientes e independientes que forman parte del estudio.

Variables dependientes

Las variables dependientes utilizadas en este análisis corresponden a la pérdida esperada en distintos segmentos del sistema crediticio mexicano, expresadas como porcentaje de los portafolios correspondientes. Separar la variable dependiente según el tipo de portafolio crediticio permite observar la heterogeneidad de riesgo en cada segmento de crédito, lo cual es fundamental tanto para mejorar la capacidad predictiva de los modelos como para hacer recomendaciones puntuales de políticas y estrategias de mitigación que sean más específicas y efectivas. Las variables dependientes son:

1. Pérdida esperada - actividad empresarial o comercial: Representa la pérdida esperada asociada a créditos otorgados para actividades empresariales o comerciales, reflejando el riesgo crediticio en el sector productivo.
2. Pérdida esperada - automotriz: Mide la pérdida esperada en el portafolio de créditos automotrices, es decir, aquellos destinados a la compra de vehículos.
3. Pérdida esperada - entidades financieras: Se refiere a la pérdida esperada en créditos otorgados a otras entidades financieras, capturando el riesgo sistémico entre intermediarios.
4. Pérdida esperada - nómina: Indica la pérdida esperada en créditos de nómina, comúnmente otorgados a trabajadores con ingresos fijos y descontados directamente de su salario.
5. Pérdida esperada - créditos personales: Representa la pérdida esperada en créditos personales, los cuales son generalmente de libre destino y sin una garantía específica.
6. Pérdida esperada - tarjeta de crédito: Mide la pérdida esperada en el portafolio de tarjetas de crédito, segmento caracterizado por alta rotación y riesgo de consumo.

Variables independientes

Por su lado, las variables independientes incluyen diversos indicadores macroeconómicos y financieros que pueden influir en el comportamiento de las pérdidas esperadas en los distintos portafolios crediticios. Estas variables son:

1. IGAE (Indicador Global de Actividad Económica): Índice mensual base 2018=100 aproxima la evolución del PIB mexicano. Se publica en valores originales y ajustados por estacionalidad, para el análisis se han considerado los valores sin ajuste por estacionalidad.
2. INPC: Índice Nacional de Precios al Consumidor (INPC) indicador que mide la variación promedio de los precios de una canasta de bienes y servicios representativa del consumo de los hogares mexicanos a lo largo del tiempo (no se usa la inflación YoY para no tener efectos base).
3. Tasa de referencia (TIIE de fondeo): es la tasa de interés interbancaria de equilibrio utilizada como referencia para el fondeo entre bancos. Su valor actual, justo por debajo de la tasa objetivo, refleja las condiciones de liquidez del sistema financiero.
4. Empleo IMSS: Número total de trabajadores definitivos registrados en el Instituto Mexicano del Seguro Social. Indicador de empleo formal.
5. Confianza empresarial: Resultados de la Encuesta Mensual de Opinión Empresarial (EMOE), mide la percepción de los empresarios sobre la situación actual y futura.
6. S&P 500: Índice bursátil que refleja el comportamiento de 500 empresas grandes en EE.UU. Promedio mensual de precios de cierre diarios.
7. IPC: El Índice de Precios y Cotizaciones (IPC) es el principal indicador de desempeño de la Bolsa Mexicana de Valores. Está compuesto por las acciones de las empresas más grandes y con mayor liquidez de México.

8. Remesas: Ingresos por remesas familiares del exterior, en millones de USD. Publicación mensual del Banco de México.
9. Saldo de la balanza comercial no petrolera: Diferencia mensual entre exportaciones e importaciones de mercancías sin considerar la balaza petrolera.
10. Agregados monetarios M1, M2,M3: M1 es el dinero en efectivo y en cuentas de cheques, es decir, el que puedes gastar al instante. M2 incluye todo lo de M1, más los ahorros y los depósitos a plazo, por lo que es un poco menos líquido, pero todavía fácil de convertir en efectivo. M3 amplía aún más e incorpora lo de M2 más valores emitidos por instituciones no bancarias, por lo que representa dinero menos accesible de inmediato.

Adicionalmente, se incorporan variables obtenidas de Google Trends que reflejan la frecuencia de búsqueda de términos como "deuda" y "empleo". La inclusión de estas variables responde a la necesidad de captar señales tempranas y complementarias sobre la percepción y las preocupaciones de la población, las cuales pueden anticipar cambios en el comportamiento crediticio antes de que estos se reflejen en los datos tradicionales. Este tipo de información, proveniente de fuentes alternativas, permite enriquecer el modelo con perspectivas que no están presentes en las bases de datos convencionales y que quizás pueden ofrecer una visión más inmediata de cambios en la confianza de los hogares o las condiciones del mercado laboral.

11. Google Trends - deuda: Esta variable representa el volumen de búsquedas realizadas en Google relacionadas con el término "deuda" en México, obtenida a partir de Google Trends. Su objetivo es capturar cambios en el interés o preocupación de la población respecto a temas de endeudamiento.
12. Google Trends - empleo: Esta variable corresponde al volumen de búsquedas en Google asociadas con el término "empleo", también basada en los datos de Google Trends para México. La inclusión de esta variable permite monitorear en tiempo real el interés de la población por cuestiones relacionadas con el empleo.

Limpieza de datos

Los datos de cartera y riesgo de crédito se tomaron de la Comisión Nacional Bancaria y de Valores (CNBV), Banxico, IMSS e INEGI, y se complementaron con datos de Google Trends (para búsquedas de "deuda" y "empleo").

Para la elaboración de la base de datos final, primero se procedió con el armado y la integración de las distintas fuentes de datos, asegurando la identificación correcta de variables numéricas, el ordenamiento de las fechas y la estandarización de formatos para garantizar la homologación entre todas las bases. Posteriormente, se trabajó en la homogeneización de los periodos disponibles, de manera que todas las series coincidieran en la misma frecuencia temporal y pudieran ser analizadas de forma conjunta. Finalmente, para preparar las variables para los análisis de series de tiempo y asegurar la estacionariedad, se aplicó la transformación de primeras diferencias utilizando un rezago mensual (lag de un mes), lo cual permitió obtener series listas para el modelado y la comparación de resultados.

Feature Engineering

Con el fin de capturar adecuadamente las dinámicas temporales inherentes a las series de tiempo financieras en los modelos de ML, se implementó lo siguiente:

- **Rezagos temporales**: Se calcularon rezagos de 1, 3 y 6 meses tanto para las variables dependientes como para las independientes, lo que permite a los modelos incorporar efectos de corto y mediano plazo.
- **Promedios móviles**: Se agregaron para suavizar fluctuaciones de alta frecuencia y mejorar la robustez del modelado.
- **Estacionalidad**: Se codificó el mes del año mediante funciones seno y coseno para capturar posibles patrones estacionales.
- **Tendencia temporal**: Se introdujo una variable de tendencia lineal creciente para capturar evolución estructural de largo plazo.

- **Manejo de valores faltantes**: Los valores faltantes generados por los rezagos fueron eliminados, asegurando que el conjunto final de entrenamiento esté completo y no contamine el modelo.

Análisis Exploratorio de datos (EDA)

Descripción general de los datos

Tabla 1: Descripción general de los datos

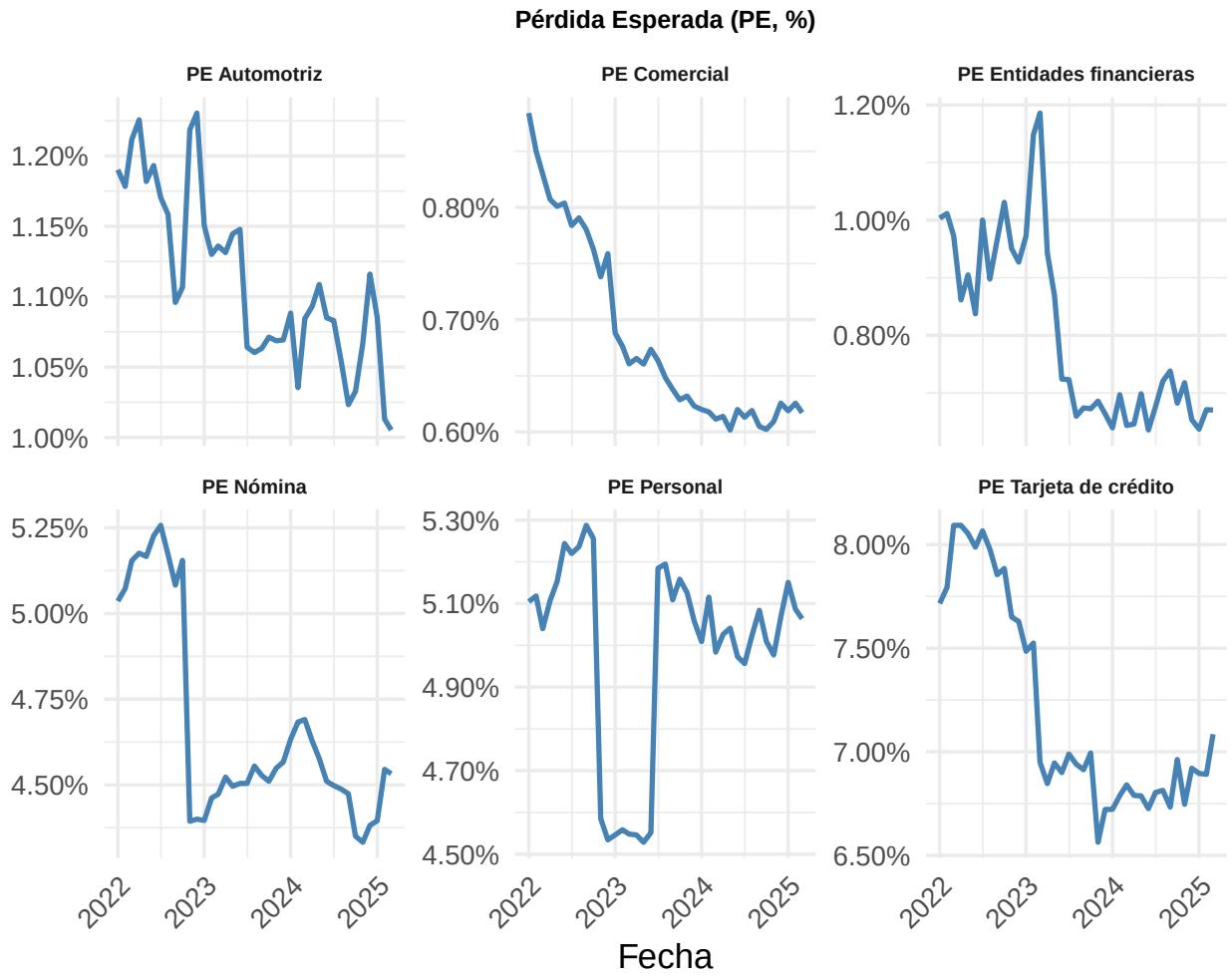
Variable	Descripción	Unidades
llave	Fecha de observación (mensual)	YYYY-MM-DD
y_pe_actividad_empresarial_o_comercial	Pérdida esperada - Actividad empresarial/comercial	Decimal (x100 = %)
y_pe_automotriz	Pérdida esperada - Crédito automotriz	Decimal (x100 = %)
y_pe_entidades_financieras	Pérdida esperada - Entidades financieras	Decimal (x100 = %)
y_pe_nomina	Pérdida esperada - Nómina	Decimal (x100 = %)
y_pe_personales	Pérdida esperada - Crédito personal	Decimal (x100 = %)
y_pe_tarjeta_de_credito	Pérdida esperada - Tarjeta de crédito	Decimal (x100 = %)
usd_mxn	Tipo de cambio USD/MXN	Pesos por USD
tiie_fondeo	Tasa de fondeo bancaria (TIIE a un día)	Decimal (x100 = %)
remesas	Remesas recibidas	Miles de millones USD
m1	Agregado monetario M1	Miles de millones MXN
m2	Agregado monetario M2	Miles de millones MXN
m3	Agregado monetario M3	Miles de millones MXN
balanza	Balanza comercial	Miles de millones USD
inpc	Índice Nacional de Precios al Consumidor	Índice
igoec	Indicador Global de la Actividad Económica (IGOE)	Índice
igae	IGAE - Base 2018 = 100	Índice
google_deuda	Tendencia de búsqueda Google: 'deuda'	Índice
google_empleo	Tendencia de búsqueda Google: 'empleo'	Índice
nacional	Puestos de trabajo registrados IMSS	Millones
s_p_bmv_ipc	Índice bursátil S&P/BMV IPC	Índice
puestos_trabajo	Puestos de trabajo registrados IMSS (copia)	Millones

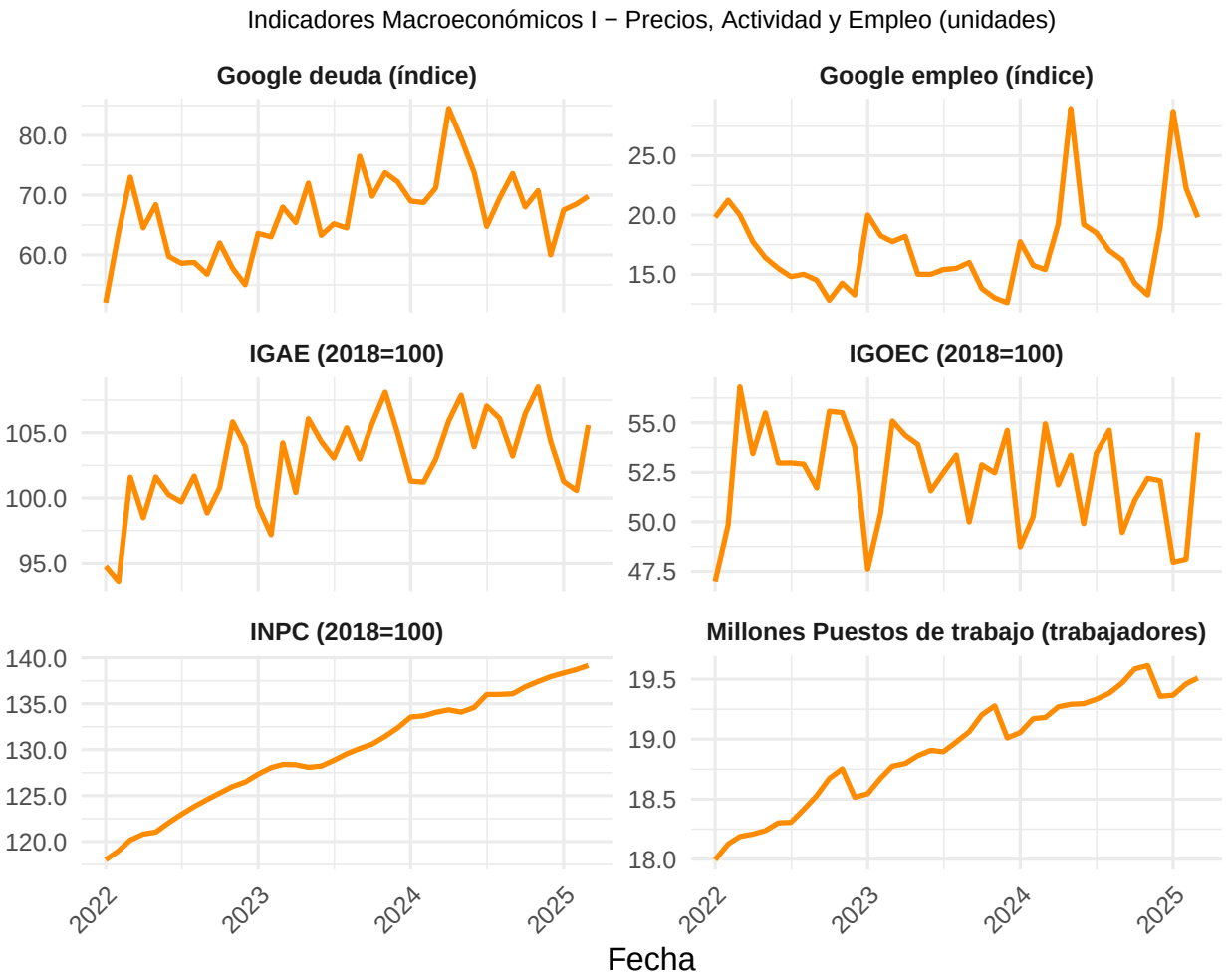
Estadística descriptiva

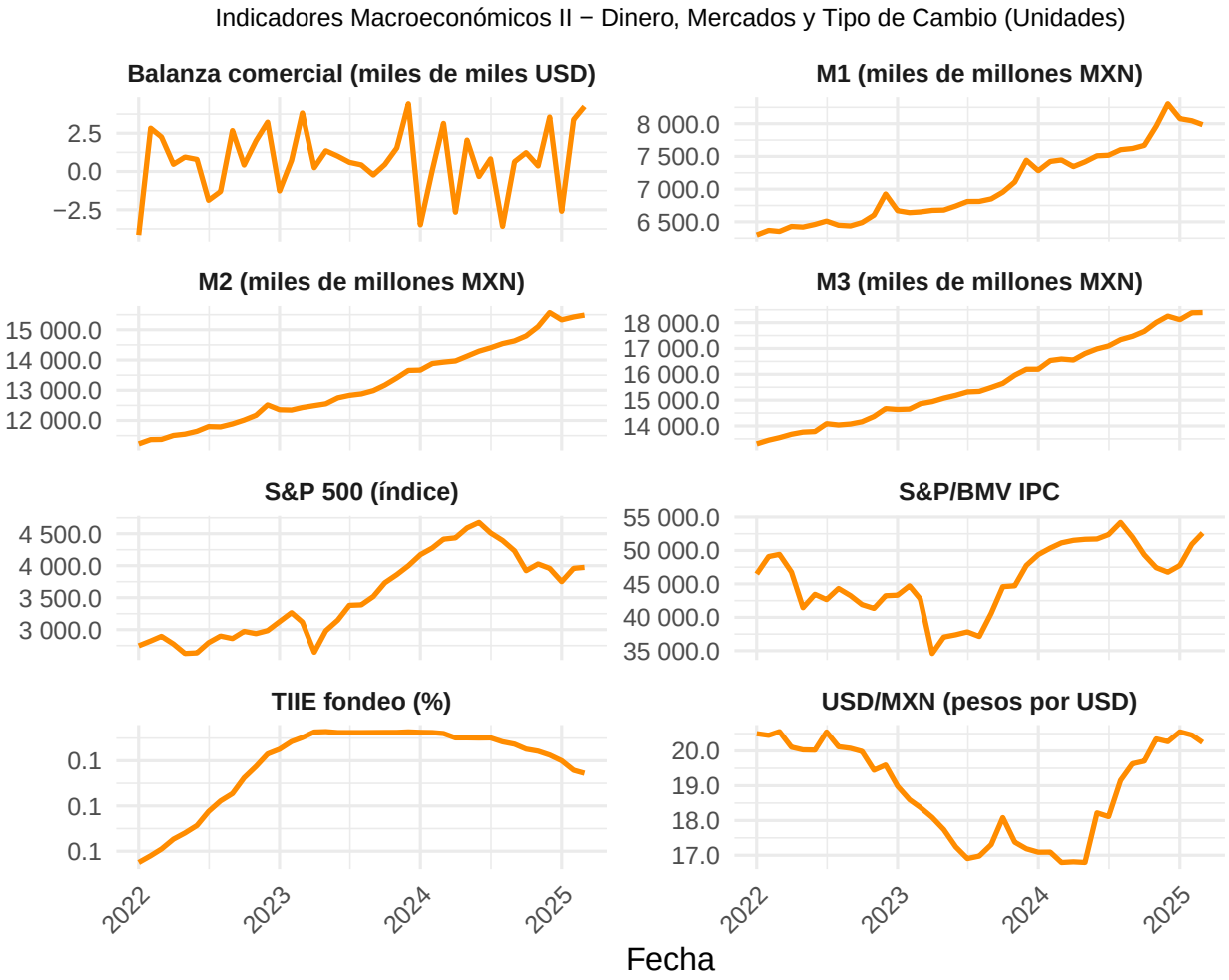
Tabla 2: Resumen estadístico de variables numéricas

Variable	N	Mean	SD	Q1	Median	Q3	Min	Max	Missing
PE_Comercial	39	0.0068	0.0008	0.0062	0.0065	0.0076	0.0060	0.0088	0
PE_Automotriz	39	0.0111	0.0006	0.0107	0.0110	0.0115	0.0101	0.0123	0
PE_EFinancieras	39	0.0081	0.0016	0.0067	0.0072	0.0095	0.0064	0.0119	0
PE_Nomina	39	0.0467	0.0030	0.0448	0.0453	0.0486	0.0433	0.0526	0
PE_Personales	39	0.0499	0.0024	0.0497	0.0506	0.0514	0.0453	0.0529	0
PE_TC	39	0.0721	0.0051	0.0681	0.0695	0.0768	0.0656	0.0809	0
USD/MXN	39	18.8600	1.3908	17.3422	19.1515	20.1148	16.7918	20.5562	0
TIEF	39	0.0987	0.0179	0.0934	0.1073	0.1124	0.0550	0.1129	0
Remesas	39	5.1586	0.5366	4.8014	5.1894	5.5764	3.9285	6.2069	0
M1	39	7050.6328	578.3790	6556.8237	6850.7445	7477.5027	6297.4821	8304.5191	0
M2	39	13175.4144	1323.7676	12091.4950	12878.4615	14209.7910	11226.0641	15570.8611	0
M3	39	15656.5281	1572.2897	14263.5408	15337.2807	16892.3191	13305.4630	18391.2069	0
Balanza	39	0.7237	2.1699	-0.0791	0.7251	2.1522	-4.1519	4.4246	0
INPC	39	129.8023	6.0551	125.6365	129.5450	134.4650	118.0020	139.1610	0
IGOE	39	52.2923	2.4729	50.3495	52.8820	54.1340	46.9940	56.8180	0
IGAE	39	102.8016	3.4964	100.6749	103.0628	105.6500	93.6058	108.5259	0
G-Deuda	39	66.8359	6.7719	63.1250	68.0000	70.9750	52.0000	84.5000	0
G-Empleo	39	17.2090	3.7108	14.9000	16.2000	19.1000	12.6000	29.0000	0
IMSS	39	18.9123	0.4611	18.5383	18.9762	19.2934	17.9943	19.6139	0
IPC	39	45765.3899	5079.9530	42693.7977	46479.2197	49885.3379	34579.4830	54198.5503	0
S&P500	39	3523.0405	655.9250	2917.6128	3387.2355	4012.9305	2624.9894	4678.1577	0

Exploración univariada



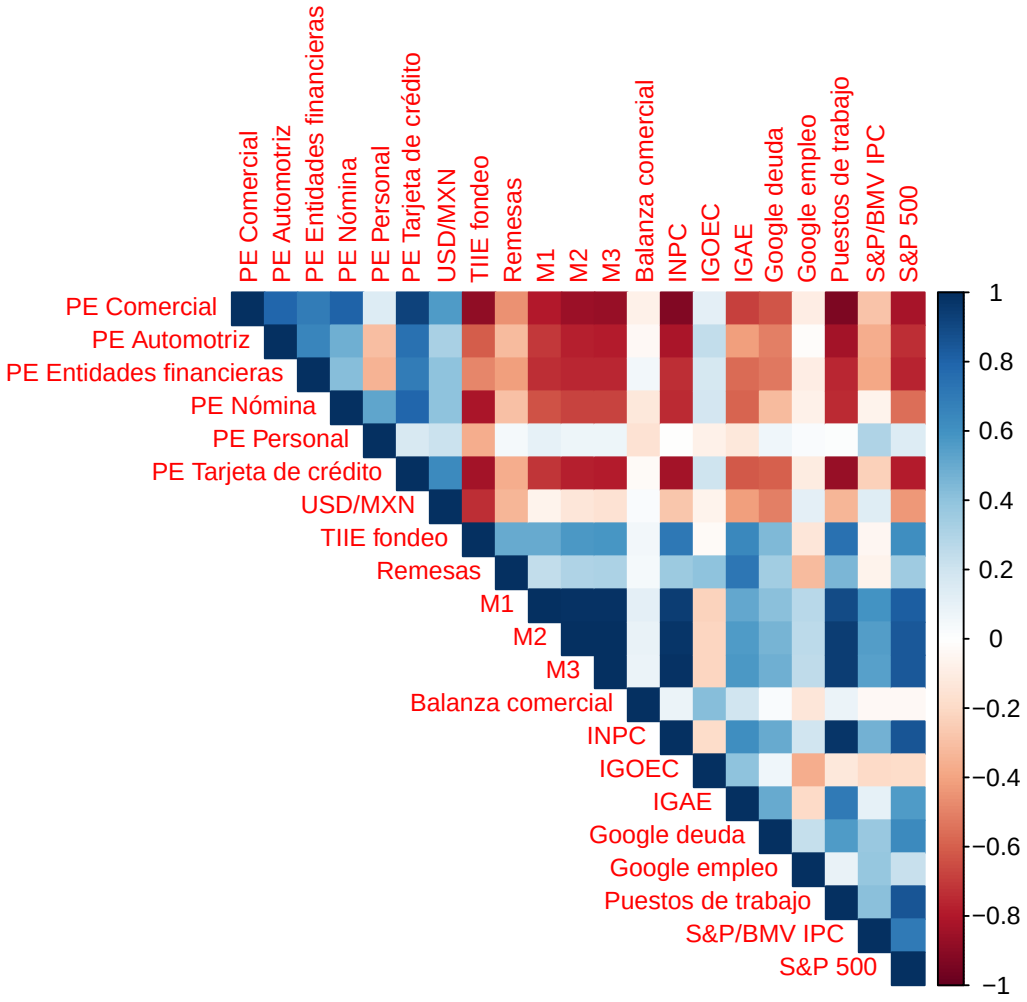

Indicadores Macroeconómicos I − Precios, Actividad y Empleo (unidades)
Google deuda (índice)
Google empleo (índice)
IGAE (2018=100)
IGOEC (2018=100)
INPC (2018=100)
Millones Puestos de trabajo (trabajadores)
80.0
70.0
60.0
25.0
20.0
15.0
105.0
100.0
95.0
55.0
52.5
50.0
47.5
140.0
135.0
130.0
125.0
120.0
19.5
19.0
18.5
18.0
2022
2023
2024
2025
Fecha



Los gráficos univariados revelan las dinámicas temporales que hay que tomar en cuenta antes de modelar las pérdidas crediticias esperadas. Los agregados monetarios (M1, M2, M3), la inflación (INPC) y el empleo (puestos de trabajo) muestran fuertes tendencias al alza, lo que sugiere un crecimiento a largo plazo y la necesidad de transformaciones como la diferenciación. Las tasas de interés (TIIE) y los tipos de cambio (USD/MXN) muestran un comportamiento cíclico alineado con los cambios recientes en la política monetaria, lo que podría tener efectos rezagados en el desempeño crediticio. Los indicadores del ciclo económico, como el IGAE y las variables de Google Trends, muestran fluctuaciones a corto plazo, lo que sugiere su potencial como predictores de alta frecuencia del estrés del prestatario.

En cuanto a las variables de pérdida crediticia, éstas presentan comportamientos diversos según el tipo de producto. Mientras que segmentos como las tarjetas de crédito y la actividad empresarial muestran descensos constantes, posiblemente debido a mejoras en la calidad de la cartera, otros, como la nómina y el personal, revelan cambios estructurales o estancamientos, lo que sugiere posibles cambios de régimen o redefiniciones contables.

Matriz de correlación



La matriz de correlación revela fuertes relaciones entre los indicadores macrofinancieros y las tasas de pérdida esperadas en diferentes productos crediticios. La mayoría de las tasas de pérdida presentan una alta correlación positiva entre sí, lo que sugiere que factores macroeconómicos sistémicos influyen simultáneamente en múltiples segmentos crediticios. Destaca que la variable *Pérdida esperada - Crédito personal* muestra correlaciones bajas o incluso negativas con otras variables de pérdida, lo que indica que podría estar impulsado por dinámicas idiosincrásicas o no observadas. Las pérdidas crediticias también presentan una correlación negativa con los agregados monetarios (M1, M2, M3), el empleo y la inflación, lo que implica que una mayor actividad económica se asocia con un menor riesgo crediticio.

Entre los predictores macroeconómicos, las tasas de interés muestran fuertes correlaciones negativas con todas las variables de pérdida crediticia por lo que tasas más altas coinciden con menores incumplimientos a corto plazo, lo que posiblemente refleja estándares de préstamo más estrictos o efectos de selección. Por el contrario, el tipo de cambio y las tendencias de búsqueda en Google presentan una correlación positiva con las pérdidas crediticias, lo que indica posibles señales de estrés. Estos hallazgos resaltan la importancia del contexto macroeconómico en la modelización del riesgo crediticio.

Pruebas de estacionariedad

- **Prueba de Raíz Unitaria de Dickey & Fuller**

La prueba de Dickey-Fuller Aumentada (DFA) se usa para detectar estadísticamente la presencia de conducta tendencial estocástica en las series temporales de las variables mediante un contraste de hipótesis. En nuestro

caso, se utilizó la prueba DFA para evaluar la presencia de raíz unitaria en las series temporales de nuestra base de datos. Los resultados indican que, en su mayoría, las series analizadas requieren una transformación mediante primeras diferencias para alcanzar la estacionariedad.

Una serie de tiempo debe ser estacionaria para que los modelos econométricos puedan aplicarse correctamente. Si una serie no es estacionaria, sus propiedades estadísticas —como la media, la varianza y la covarianza— cambian con el tiempo, lo cual invalida los supuestos fundamentales de los modelos clásicos. Esto impide usar procedimientos estándar de inferencia estadística y genera relaciones engañosas entre las variables. Como afirma Walter Enders:

> "If a time series is nonstationary, it is not possible to model it in terms of a finite number of parameters that summarize its behavior over time. The reason is that the process generating the series is not stable over time. Consequently, any coefficients obtained from fitting a model to the data are not likely to be constant over time" — Walter Enders, Applied Econometric Time Series, 4th edition, Chapter 2.3, p. 50 .

Esto significa que los coeficientes estimados pueden no ser válidos en otros periodos, los errores se acumulan y las predicciones pierden validez. Por ello, es indispensable asegurar que las series sean estacionarias antes de modelarlas.

Resultados del Test ADF por Variable

Evaluación de Estacionariedad con tseries::adf.test (k = 0)

variable	adf_stat	adf_pvalue	estacionaria	usable
Puestos de trabajo	−3.040	0.166	FALSE	NO
M1	−2.751	0.279	FALSE	NO
INPC	−2.645	0.321	FALSE	NO
PE Entidades financieras	−2.339	0.440	FALSE	NO
M2	−2.290	0.460	FALSE	NO
M3	−2.252	0.475	FALSE	NO
PE Personal	−2.129	0.523	FALSE	NO
PE Nómina	−1.853	0.631	FALSE	NO
S&P/BMV IPC	−1.804	0.650	FALSE	NO
PE Comercial	−1.690	0.695	FALSE	NO
PE Tarjeta de crédito	−1.189	0.891	FALSE	NO
S&P 500	−1.145	0.903	FALSE	NO
USD/MXN	−0.538	0.975	FALSE	NO
TIIE fondeo	−0.342	0.984	FALSE	NO
Balanza comercial	−8.164	0.010	TRUE	OK
IGOEC	−6.230	0.010	TRUE	OK
IGAE	−4.811	0.010	TRUE	OK
Google deuda	−4.045	0.018	TRUE	OK
Remesas	−3.631	0.044	TRUE	OK
Google empleo	−3.614	0.045	TRUE	OK
PE Automotriz	−3.614	0.045	TRUE	OK

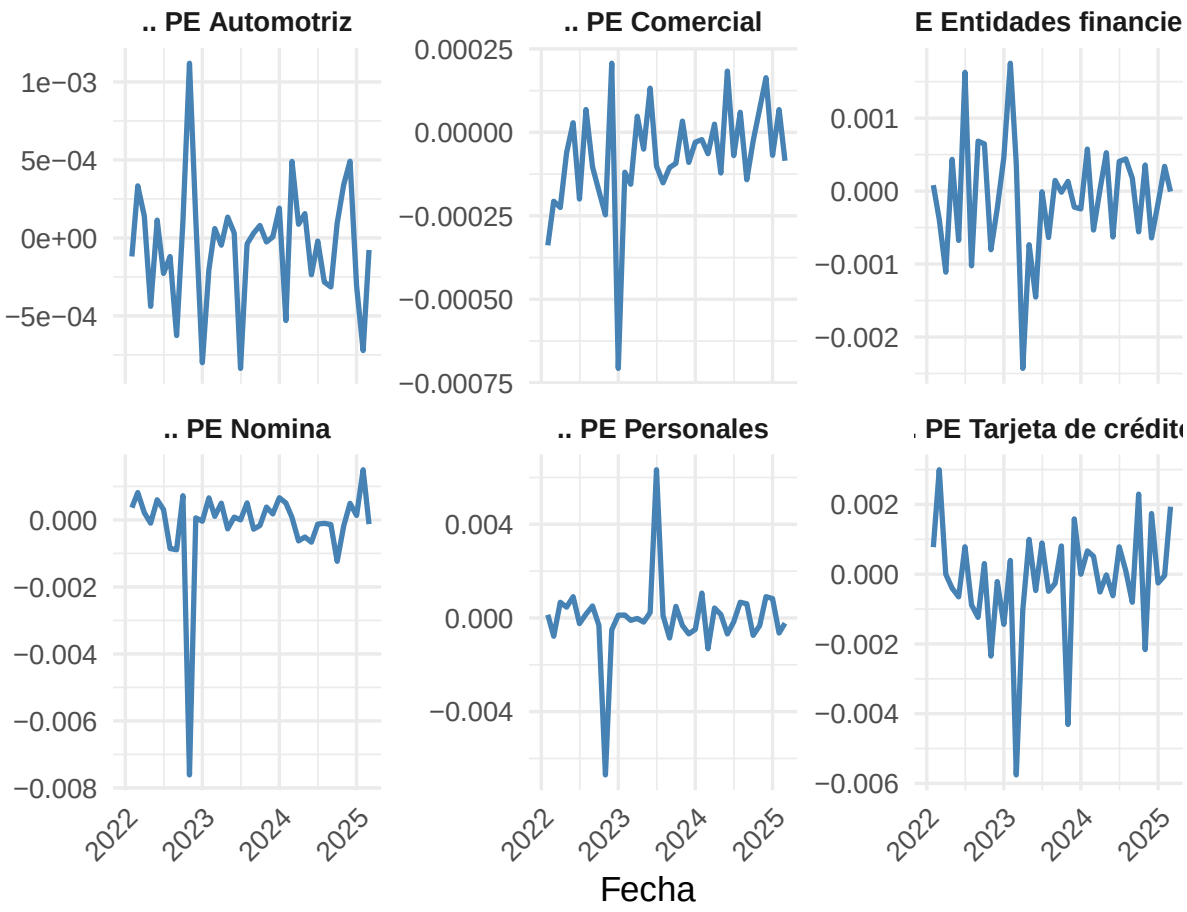
Resultados del Test ADF – Primeras Diferencias

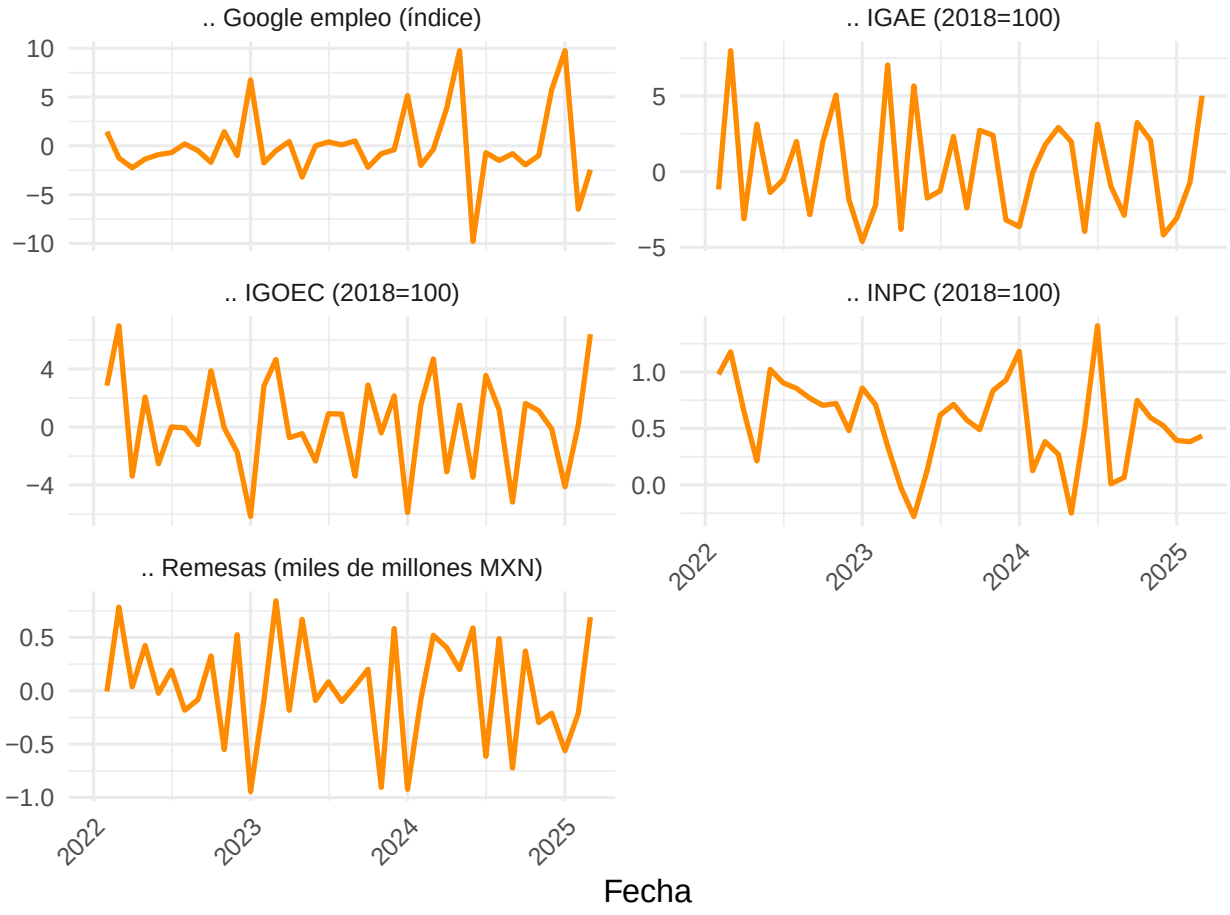
Evaluación de Estacionariedad con tseries::adf.test (k = 0)

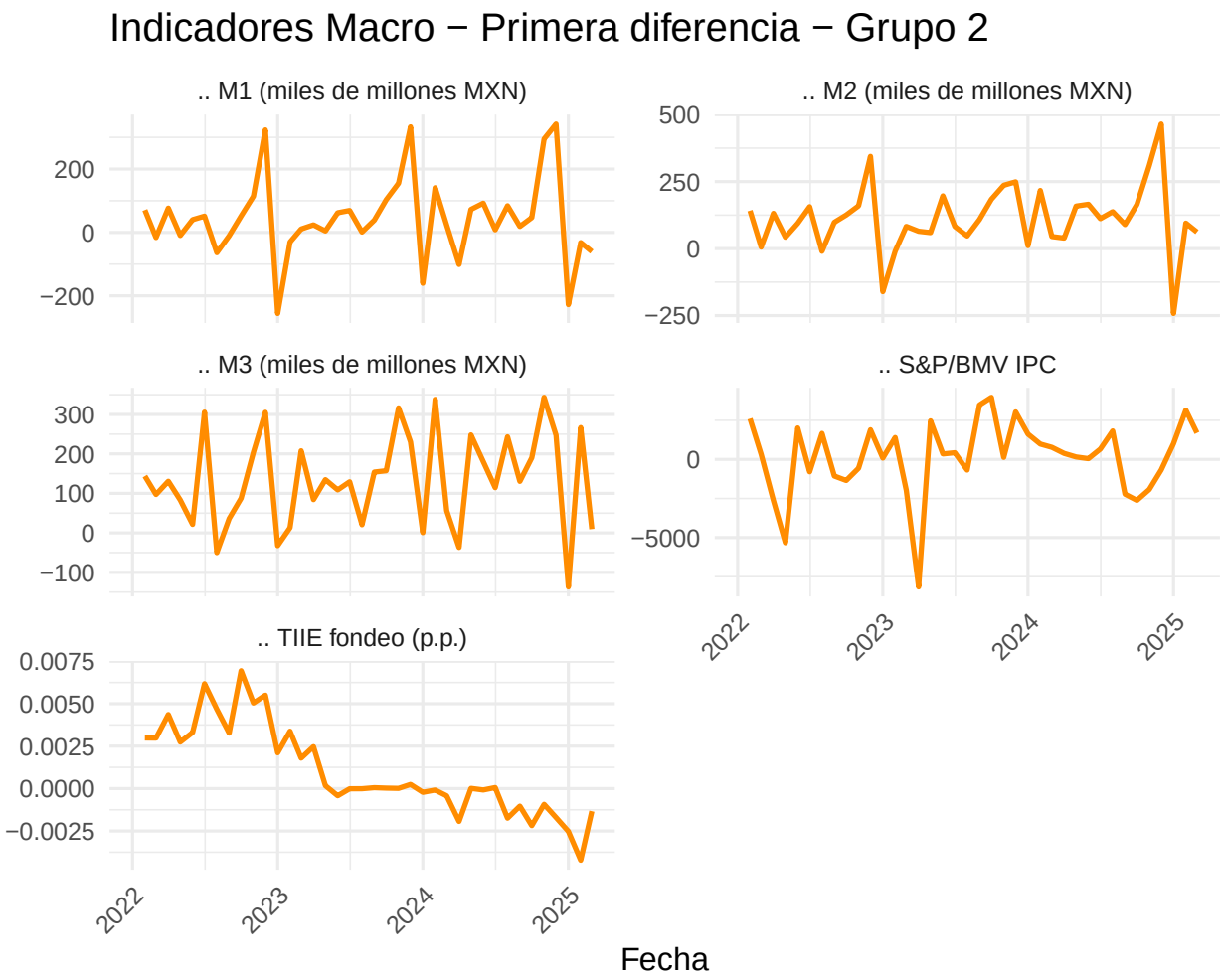
variable	adf_stat	adf_pvalue	estacionaria	usable
diff_PE Comercial	−8.41468	0.01000	TRUE	OK
diff_PE Automotriz	−5.63978	0.01000	TRUE	OK
diff_PE Entidades financieras	−6.41906	0.01000	TRUE	OK
diff_PE Nómina	−6.34084	0.01000	TRUE	OK

diff_PE Personal	−5.56040	0.01000	TRUE	OK
diff_PE Tarjeta de crédito	−7.29581	0.01000	TRUE	OK
diff_USD/MXN	−6.03114	0.01000	TRUE	OK
diff_Remesas	−9.05683	0.01000	TRUE	OK
diff_M1	−6.44392	0.01000	TRUE	OK
diff_M2	−6.75203	0.01000	TRUE	OK
diff_M3	−7.65344	0.01000	TRUE	OK
diff_Balanza comercial	−11.08885	0.01000	TRUE	OK
diff_INPC	−4.61570	0.01000	TRUE	OK
diff_IGOEC	−7.27170	0.01000	TRUE	OK
diff_IGAE	−8.05037	0.01000	TRUE	OK
diff_Google deuda	−8.25949	0.01000	TRUE	OK
diff_Google empleo	−6.76881	0.01000	TRUE	OK
diff_Puestos de trabajo	−5.70349	0.01000	TRUE	OK
diff_S&P/BMV IPC	−5.27137	0.01000	TRUE	OK
diff_S&P 500	−5.04541	0.01000	TRUE	OK
diff_TIIE fondeo	−4.14342	0.01483	TRUE	OK

Tras aplicar la transformación de primeras diferencias a todas las series temporales, se comprobó mediante el Test ADF que las series transformadas cumplen con el criterio de estacionariedad (p-valor < 0.05 en todos los casos). Por tanto, las primeras diferencias de las variables pueden utilizarse de forma adecuada en los modelos de predicción.


Primeras diferencias: Series con 'y'
.. PE Automotriz
.. PE Comercial
E Entidades financie
.. PE Nomina
.. PE Personales
PE Tarjeta de crédit
1e−03
5e−04
0e+00
−5e−04
0.00025
0.00000
−0.00025
−0.00050
−0.00075
0.001
0.000
−0.001
−0.002
0.000
−0.002
−0.004
−0.006
−0.008
0.004
0.000
−0.004
0.002
0.000
−0.002
−0.004
−0.006
2022
2023
2024
2025
Fecha


Indicadores Macro − Primera diferencia − Grupo 1
.. Google empleo (índice)
.. IGAE (2018=100)
.. IGOEC (2018=100)
.. INPC (2018=100)
.. Remesas (miles de millones MXN)
10
5
0
−5
−10
5
0
−5
4
0
−4
1.0
0.5
0.0
0.5
0.0
−0.5
−1.0
2022
2023
2024
2025
Fecha



Modelado

Para el análisis de nuestras variables, planeamos hacer la comparación entre los siguientes modelos:

1. **ARIMA**
   - Objetivo: Modelar y predecir la dinámica interna de una sola serie temporal de pérdida esperada, sirviendo como línea base clásica.
   - Ventajas: Capta patrones autorregresivos, tendencias y ciclos; es ampliamente aceptado y fácil de interpretar.
   - Consideraciones: La serie debe ser estacionaria; seleccionar correctamente los parámetros (p, d, q); revisar residuos.
2. **SARIMA**
   - Objetivo: Incorporar patrones estacionales en series de tiempo, útil si las pérdidas presentan estacionalidad o ciclos repetitivos.
   - Ventajas: Captura tanto dinámica general como efectos estacionales; permite predecir en presencia de ciclos regulares.
   - Consideraciones: Identificar y ajustar adecuadamente componentes estacionales (P, D, Q, s); asegurar estacionariedad.
3. **Métodos basados en árboles**
   - Objetivo: Detectar patrones complejos y no lineales en el riesgo crediticio, integrando variables macroeconómicas, financieras y alternativas (como Google Trends).
   - Ventajas: Alta precisión, maneja grandes volúmenes de datos, robusto a variables irrelevantes y

permite interpretar importancia de variables.
- Consideraciones: Evitar fuga de información temporal (data leakage), usar variables rezagadas, validar con divisiones temporales adecuadas (walk-forward).

SARIMA

Usamos modelos SARIMA para pronosticar la perdida esperada de portafolio de crédito, como actividad empresarial, nómina y tarjeta. Estos modelos permiten capturar patrones en el tiempo, incluyendo tendencias y cierta estacionalidad. Ajustamos automáticamente, con el uso de autoarima, el modelo más adecuado para cada serie y generamos predicciones a corto plazo. Aunque el ajuste en el periodo histórico fue aceptable, las proyecciones muestran limitaciones y no logran reflejar por completo los ciclos estacionales observados en los datos.

Para el pronóstico de la **pérdida esperada del portafolio de crédito** se utiliza la La forma funcional general de un modelo SARIMA $(p,d,q)(P,D,Q)_s$ se expresa como:

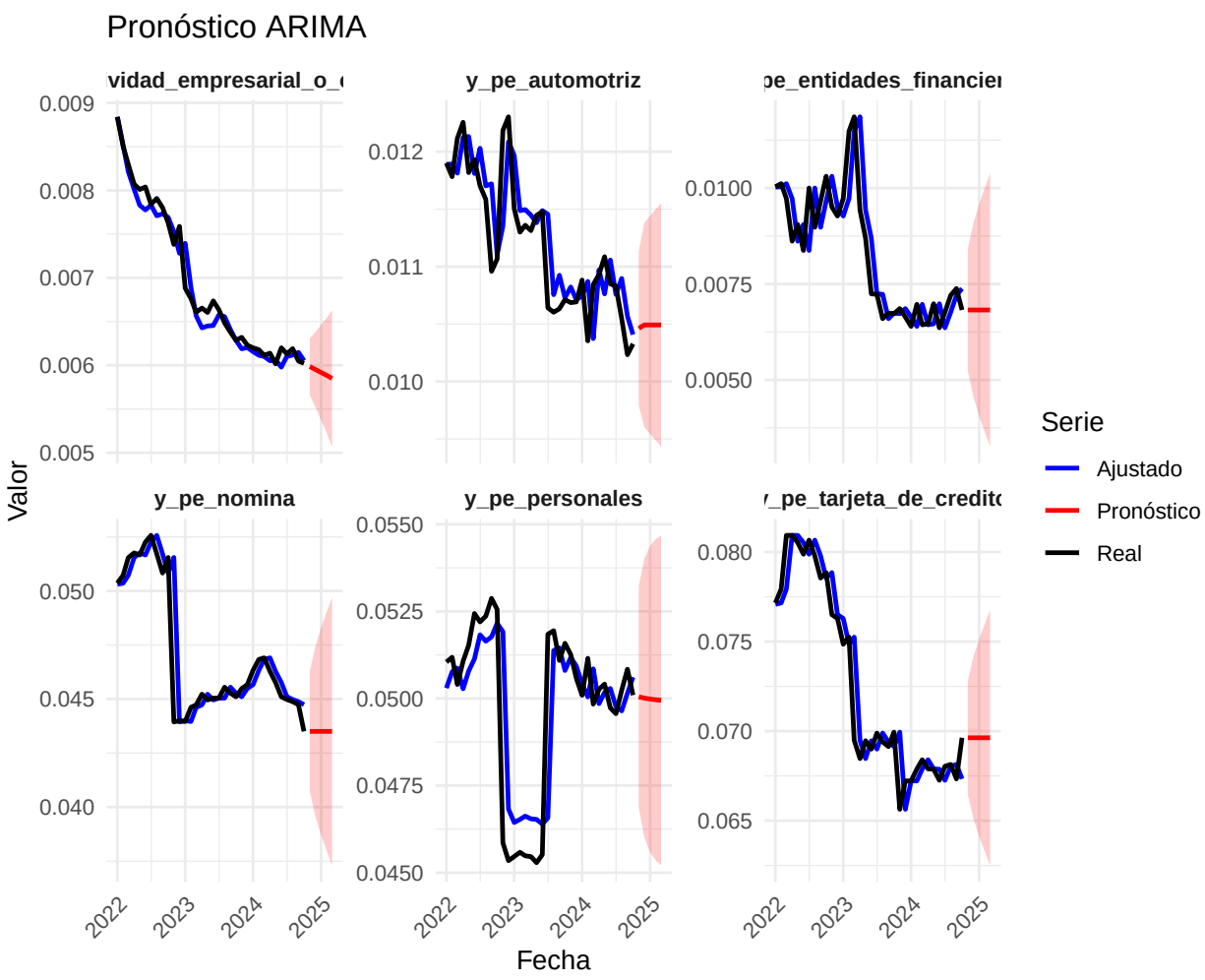
$$\Phi_P(L^s)\,\phi_p(L)(1-L)^d(1-L^s)^D y_t = \Theta_Q(L^s)\,\theta_q(L)\,\varepsilon_t$$

donde:

- p, d, q representan el orden del modelo ARIMA (autoregresivo, diferencia y promedio móvil),
- (P, D, Q) representan el orden de los componentes estacionales,
- (s) es la periodicidad de la estacionalidad (por ejemplo, 12 para datos mensuales),
- $(phi_p(L)$ y $(Phi_P(L^s)$ son los polinomios autorregresivos,
- $(theta_q(L)$ y $(\Theta_Q(L^s)$ son los polinomios de promedio móvil,
- (L) es el operador de rezago,
- $(varepsilon_t)$ es el término de error en el periodo \(t\).

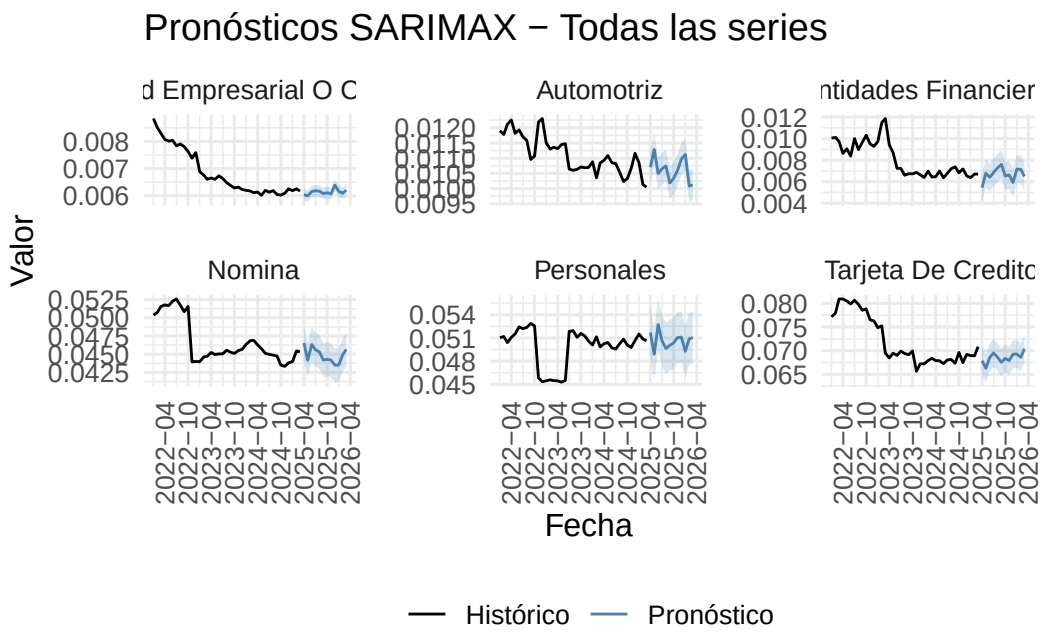
A continuación se muestra el resultado del modelo ajustado automáticamente (usando 'auto.arima') y las predicciones a corto plazo:

```
Modelo para y_pe_actividad_empresarial_o_comercial: ARIMA(1,2,1)Modelo para y_pe_automotriz: ARIMA(0,1,2
```



SARIMAX

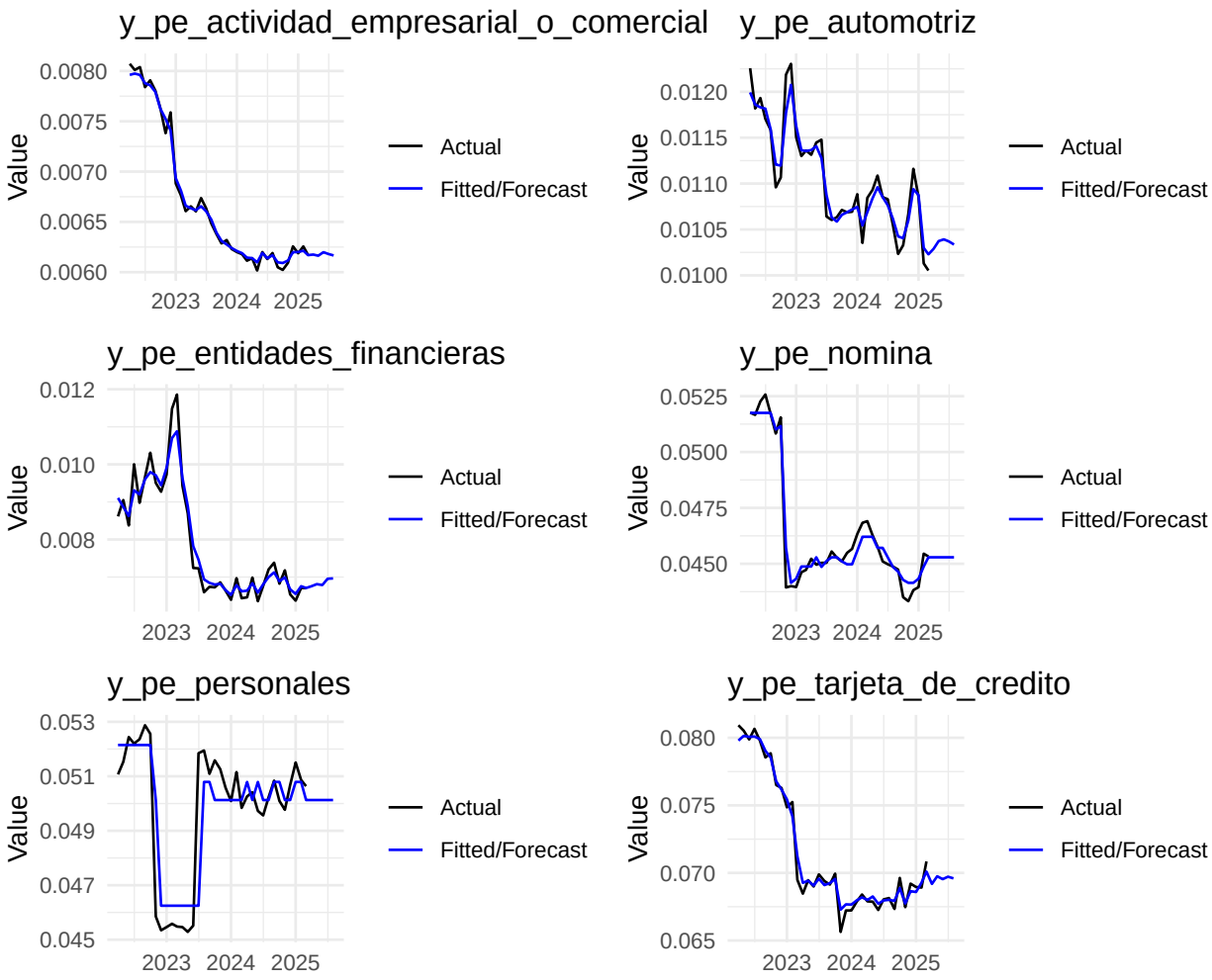
A diferencia del SARIMA, el modelo SARIMAX incorpora variables externas (en este caso hemos ocupado todas nuestras variables exogenas disponibles) que influyen en el comportamiento de los créditos. Al integrarlas, las proyecciones muestran mayor sensibilidad a los cambios económicos y capturan mejor la dinámica reciente. En general, las predicciones de SARIMAX resultaron más alineadas con los patrones observados, especialmente en series con mayor variabilidad.



Árboles

Tabla 3: Resultado de los modelos

Target_Variable	Tree	Bagging	RandomForest	XGBoost
y_pe_actividad_empresarial_o_comercial	0.0007026	0.0003579	0.0003103	0.0007941
y_pe_automotriz	0.0008012	0.0004847	0.0004949	0.0006879
y_pe_entidades_financieras	0.0013684	0.0003667	0.0003518	0.0003552
y_pe_nomina	0.0026400	0.0017604	0.0016970	0.0015211
y_pe_personales	0.0011138	0.0022471	0.0015941	0.0019321
y_pe_tarjeta_de_credito	0.0036495	0.0010725	0.0012266	0.0013376



Dado que los modelos basados en árboles no consideran inherentemente patrones temporales, diseñamos características adicionales para codificar explícitamente la dinámica temporal. Estas incluyeron rezagos de 1, 3 y 6 meses tanto para los objetivos como para los predictores, promedios móviles para suavizar la volatilidad local, indicadores estacionales (mes del año, estacionalidad seno/coseno) y una variable de tendencia temporal lineal. Tras eliminar los valores faltantes generados por el rezago, construimos un conjunto de datos de entrenamiento para el aprendizaje supervisado.

Implementamos y comparamos cuatro modelos para cada variable objetivo: un árbol de decisión podado, bagging (mediante bosques aleatorios con muestreo completo de variables), un bosque aleatorio estándar y XGBoost.

Árboles de Decisión para Regresión

La predicción de un árbol de decisión se basa en dividir el espacio de las variables independientes en regiones disjuntas $R_1, R_2, \ldots, R_M$, y asignar a cada región una predicción constante igual al promedio de los valores de la variable objetivo en dicha región:

$$\hat{f}(x) = \frac{1}{|R_l|} \sum_{i \in R_l} y_i$$

La partición se construye recursivamente, seleccionando en cada paso la variable y punto de corte que minimizan el error cuadrático dentro de las particiones resultantes. Esto se formaliza con el criterio de suma residual de cuadrados (RSS):

$$RSS = \sum_{i \in R_1} (y_i - \bar{y}{R_1}^2 + \sum i \in R_2 (y_i - \bar{y}_{R_2})^2$$

Random Forest

Random Forest es un método de ensamblado (ensemble) que promedia múltiples árboles de decisión, entrenados sobre muestras bootstrap del conjunto de datos y con una selección aleatoria de variables en cada división. La predicción final es:

$$\hat{f}_{RF}(x) = \frac{1}{B} \sum b = 1^B \hat{f}^{(b)}(x)$$

Este método reduce la varianza del modelo sin aumentar sustancialmente el sesgo, mejorando así su capacidad de generalización.

Gradient Boosting y XG Boost

Gradient Boosting entrena secuencialmente árboles que corrigen los errores del modelo ensamblado hasta ese punto:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

$F_m(x)$: modelo total después de m interaciones,

$h_m(x)$: nuevo árbol que ajusta los residuos,

η: tasa de aprendizaje o shrinkage

XGBoost, una versión optimizada de boosting, agrega regularización explícita para controlar la complejidad del modelo. La función objetivo general es:

$$\mathcal{L}t = \sum_{i=1}^{n} l(y_i, \hat{y}i(t)) + \sum k = 1^T \left(\gamma + \frac{1}{2} \lambda w_k^2 \right)$$

$l(\cdot)$: función de pérdida (ej. error cuadrático),

T: número de hojas en el árbol,

w_k: peso de la k-ésima hoja,

γ, λ: hiperparámetros de regularización.

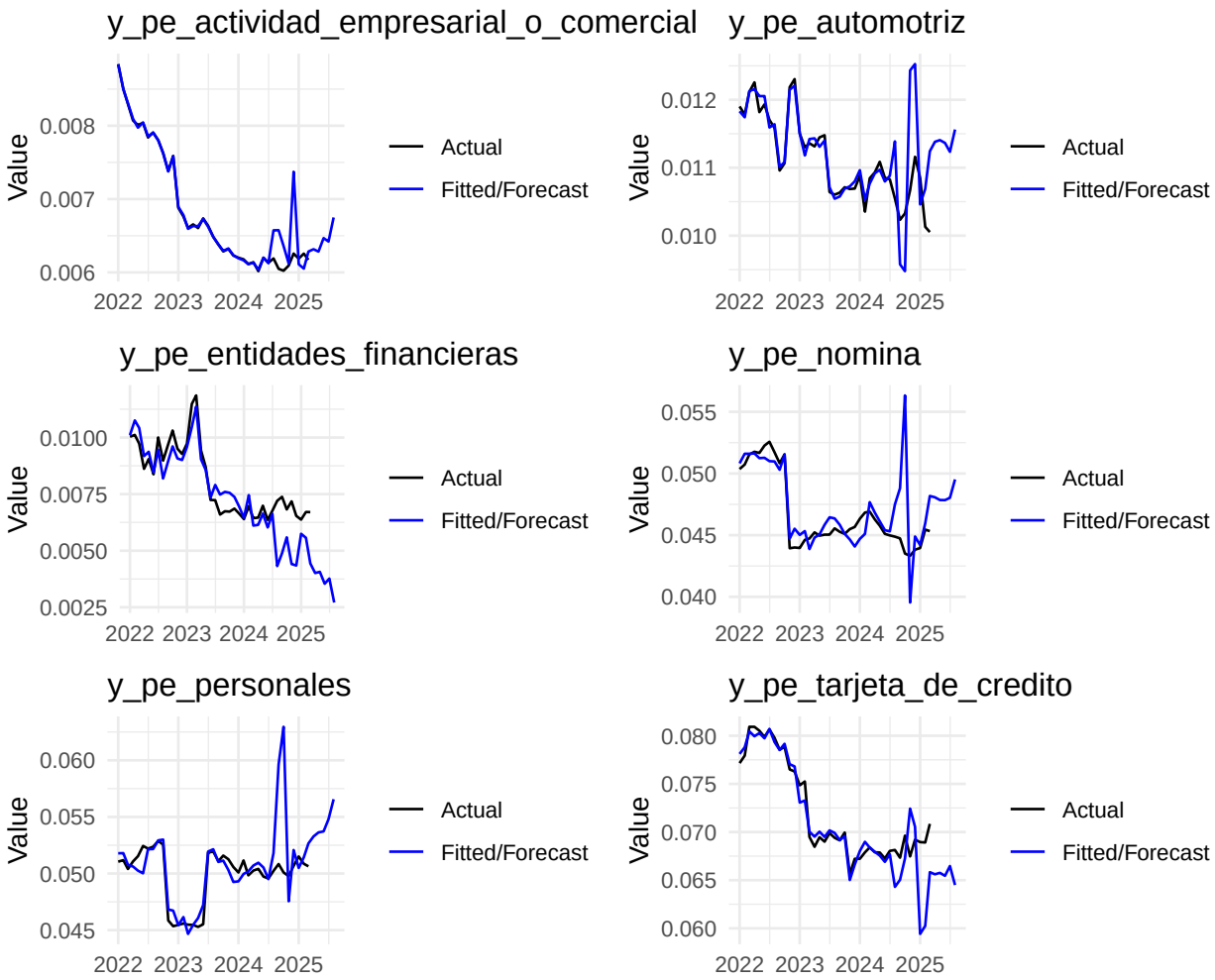
Cada modelo se entrenó con el primer 80 % de los datos ordenados temporalmente y se evaluó con el 20 % restante, utilizando RMSE como métrica de rendimiento. Los resultados mostraron que los métodos de conjunto, en particular los bosques aleatorios y XGBoost, superaron consistentemente a los árboles de decisión individuales, reduciendo considerablemente el error de pronóstico.

Una vez identificado el mejor modelo para cada objetivo, lo reentrenamos utilizando toda la muestra histórica y generamos pronósticos con cinco meses de antelación.

Prophet

Tabla 4: Comparación

Target_Variable	Model	RMSE_tree	RMSE_prophet
y_pe_actividad_empresarial_o_comercial	RandomForest	0.0003103	0.0004800
y_pe_automotriz	Bagging	0.0004847	0.0010435
y_pe_entidades_financieras	RandomForest	0.0003518	0.0021025
y_pe_nomina	XGBoost	0.0015211	0.0051460
y_pe_personales	Tree	0.0011138	0.0056752
y_pe_tarjeta_de_credito	Bagging	0.0010725	0.0055139



Prophet es un modelo estructural de series temporales diseñado para gestionar automáticamente tendencias, estacionalidad y puntos de cambio. A partir de nuestro conjunto de datos dividimos el 80 % de las observaciones para el ajuste del modelo, y el 20 % restante para la evaluación fuera de la muestra. Para cada serie, ajustamos un modelo Prophet de estacionalidad multiplicativa en la ventana de entrenamiento y, a continuación, generamos predicciones para esa ventana (valores ajustados dentro de la muestra) y para el período de prueba mantenido, más cinco meses adicionales de pronóstico real.

Prophet destaca donde predominan los patrones estructurales: su detección de puntos de cambio se adapta rápidamente a los cambios de tendencia, y la estacionalidad multiplicativa escala de forma natural las fluctuaciones estacionales a medida que aumenta la magnitud de la serie. Normalmente, igualó o superó

a nuestros modelos de árbol en segmentos con una estacionalidad fuerte y regular o cambios de tendencia pronunciados, como las pérdidas por préstamos de nómina que aumentan cíclicamente. Sin embargo, cuando los efectos autorregresivos de corto retardo (por ejemplo, las pérdidas del mes inmediatamente anterior) fueron el factor principal, nuestros conjuntos de Bosque Aleatorio y XGBoost a menudo superaron a Prophet al aprovechar directamente esas características de retardo diseñadas.

VAR

El modelo VAR (Vector Autoregresivo) es una herramienta econométrica diseñada para capturar las interdependencias dinámicas entre múltiples series temporales endógenas. A diferencia de los modelos univariados como SARIMA o SARIMAX, el VAR permite que cada variable sea explicada por sus propios rezagos y por los rezagos de las demás variables del sistema. Como complemento a los modelos individuales previamente estimados, se implementó un VAR en primeras diferencias para analizar el comportamiento conjunto de las series. Este modelo logró proyectar trayectorias con ciclos coherentes y consistentes con las dinámicas observadas en el pasado reciente, capturando adecuadamente la comovilidad entre las series consideradas.

Para el análisis se emplea un modelo de **Vectores Autorregresivos (VAR)** con rezagos \(p = 4 \), el cual permite modelar la dinámica conjunta de las variables endógenas seleccionadas (todas aquellas cuyo nombre comienza con 'y_'). En términos funcionales, el modelo puede expresarse como:

$$\mathbf{y}_t = \mathbf{c} + A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + A_3\mathbf{y}_{t-3} + A_4\mathbf{y}_{t-4} + \mathbf{u}_t$$

donde:

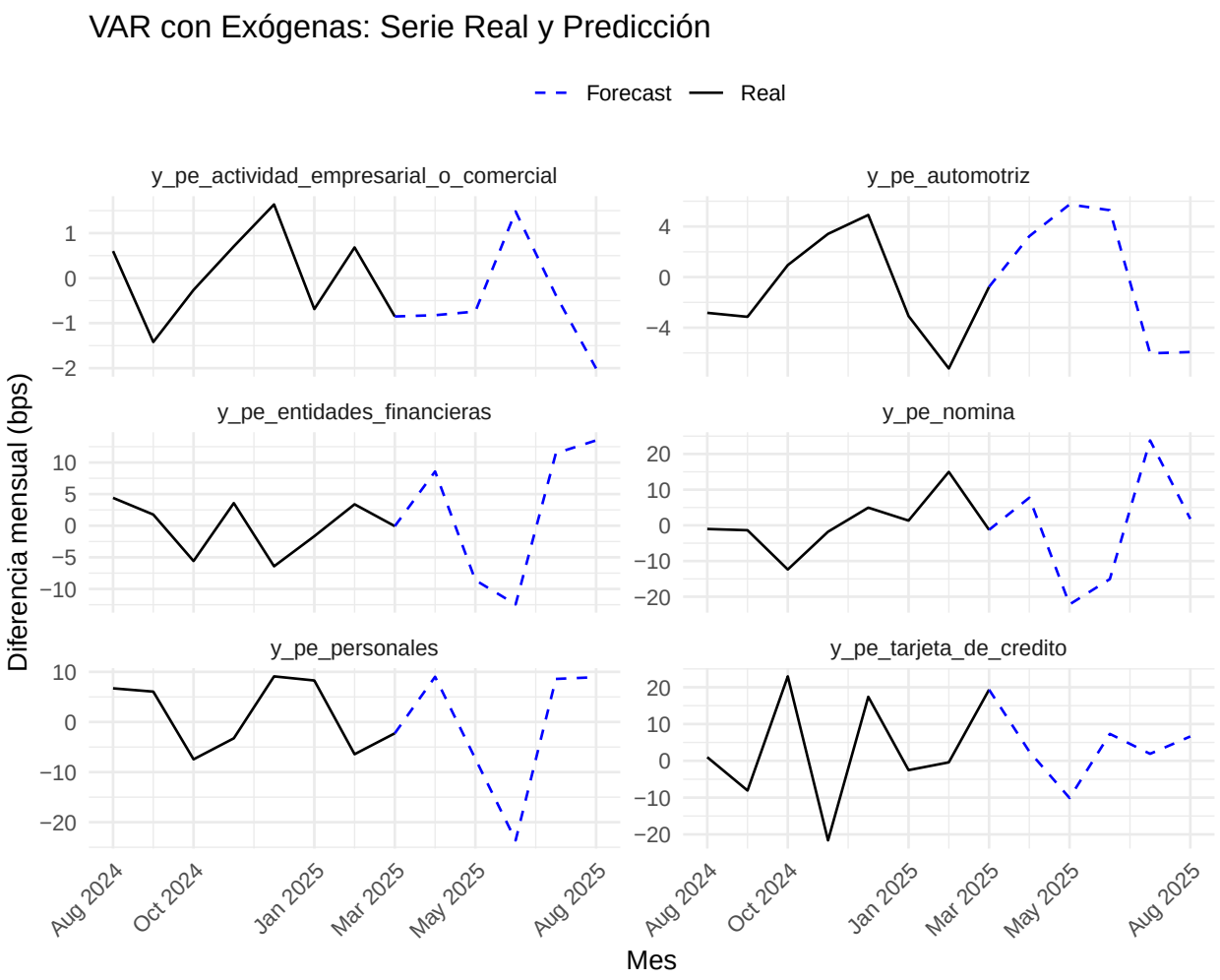
- $\mathbf{y}_t$ es el vector de variables endógenas en el tiempo \(t \),

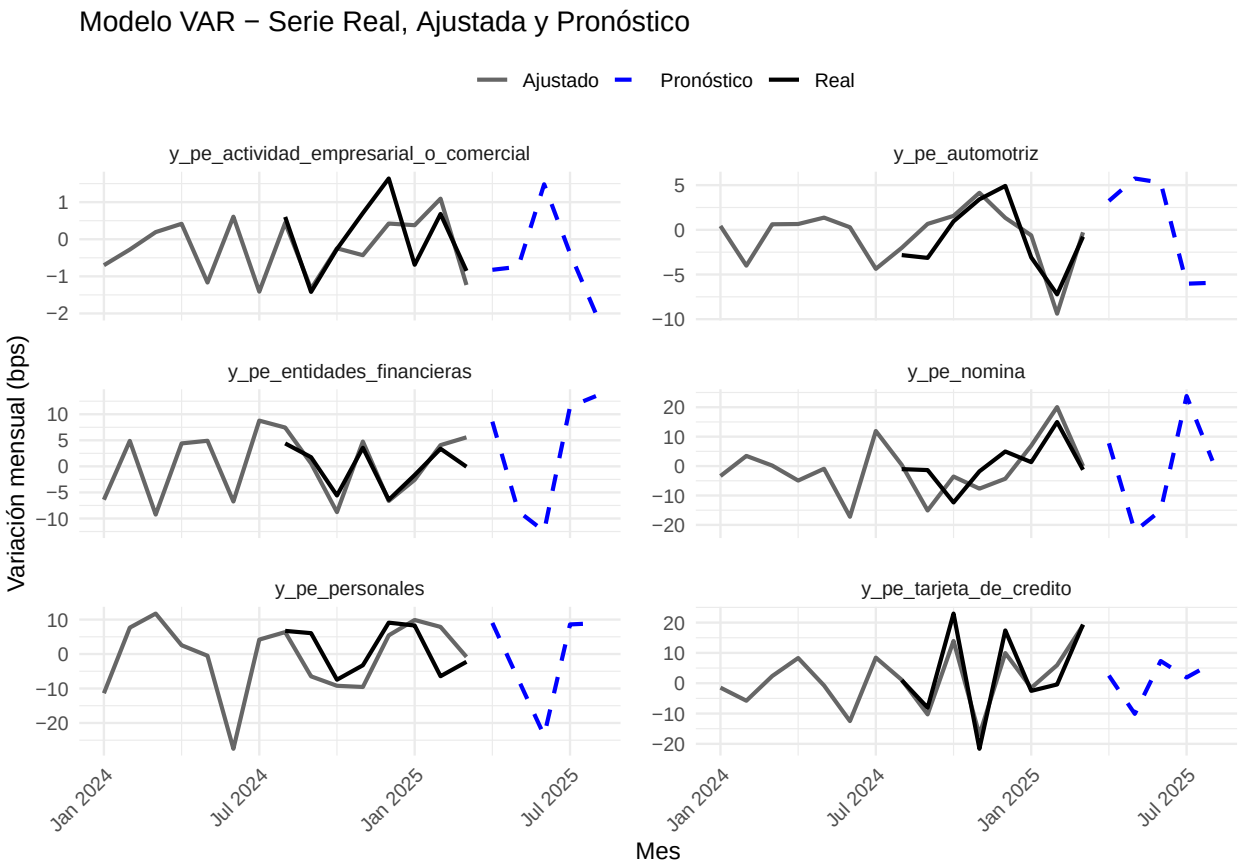
- $\mathbf{c}$ es el vector de constantes,

- A_i son las matrices de coeficientes para cada rezago \(i \),

- $\mathbf{u}_t$ es el vector de perturbaciones o errores.

A continuación se muestra el el resultado ajustado por el modelo y las predicciones:


VAR con Exógenas: Serie Real y Predicción
Forecast
Real
y_pe_actividad_empresarial_o_comercial
y_pe_automotriz
y_pe_entidades_financieras
y_pe_nomina
y_pe_personales
y_pe_tarjeta_de_credito
Diferencia mensual (bps)
Aug 2024
Oct 2024
Jan 2025
Mar 2025
May 2025
Aug 2025
Mes


Modelo VAR − Serie Real, Ajustada y Pronóstico
Ajustado
Pronóstico
Real
y_pe_actividad_empresarial_o_comercial
y_pe_automotriz
y_pe_entidades_financieras
y_pe_nomina
y_pe_personales
y_pe_tarjeta_de_credito
Variación mensual (bps)
Jan 2024
Jul 2024
Jan 2025
Jul 2025
Mes