Precios al consumidor

Manuel Toledo y Lucas Pescetto

2023-07-09

Introducción

Este proyecto surge a partir de la información provista por el SIPC (Sistema de Información de Precios al Consumidor) del Ministerio de Economía y Finanzas. Este organismo brinda información acerca de los precios de una serie de productos a través del tiempo y para distintos establecimientos en todos los departamentos de Uruguay.

Por otro lado, a raíz de la amplia diversidad de los productos de los cuales se tienen datos, se decidió tomar solamente aquellos que son parte de la CBA (Canasta Básica de Alimentos) de Uruguay.

El objetivo del trabajo es generar un análisis y visualizaciones que permitan ver las variaciones en los precios de dichos productos a lo largo del tiempo, en distintos lugares dentro del país y en establecimientos dentro de Montevideo. Esto resultaría de utilidad para ayudar a los consumidores a tomar mejores decisiones financieras a la hora de comprar alimentos.

Un objetivo adicional es el de utilizar modelos basados en series temporales para predecir los precios de los productos en el futuro.

El producto final que se busca proveer es una aplicación interactiva que para cada uno de los productos presentes en la CBA, despliegue una serie de visualizaciones descriptivas de su precio, en función de los aspectos mencionados anteriormente.

Datos

El SIPC presenta los datos en tres datasets:

- Establecimientos: es una lista de los establecimientos de los cuales se obtienen los precios. Se obtiene en la web de Catálogo abierto de datos.
- Productos: es una lista de los productos de los cuales se tienen los precios. Se obtiene en la web de Catálogo abierto de datos.
- Precios: contiene la información acerca de los precios registrados en cada momento del tiempo, para cada producto en cada establecimiento. Si bien se puede obtener en la web de Catálogo abierto de datos, debido a su tamaño se extrae con una consulta SQL.

A continuación se muestra una tabla con las variables de cada dataset:

(tablas)

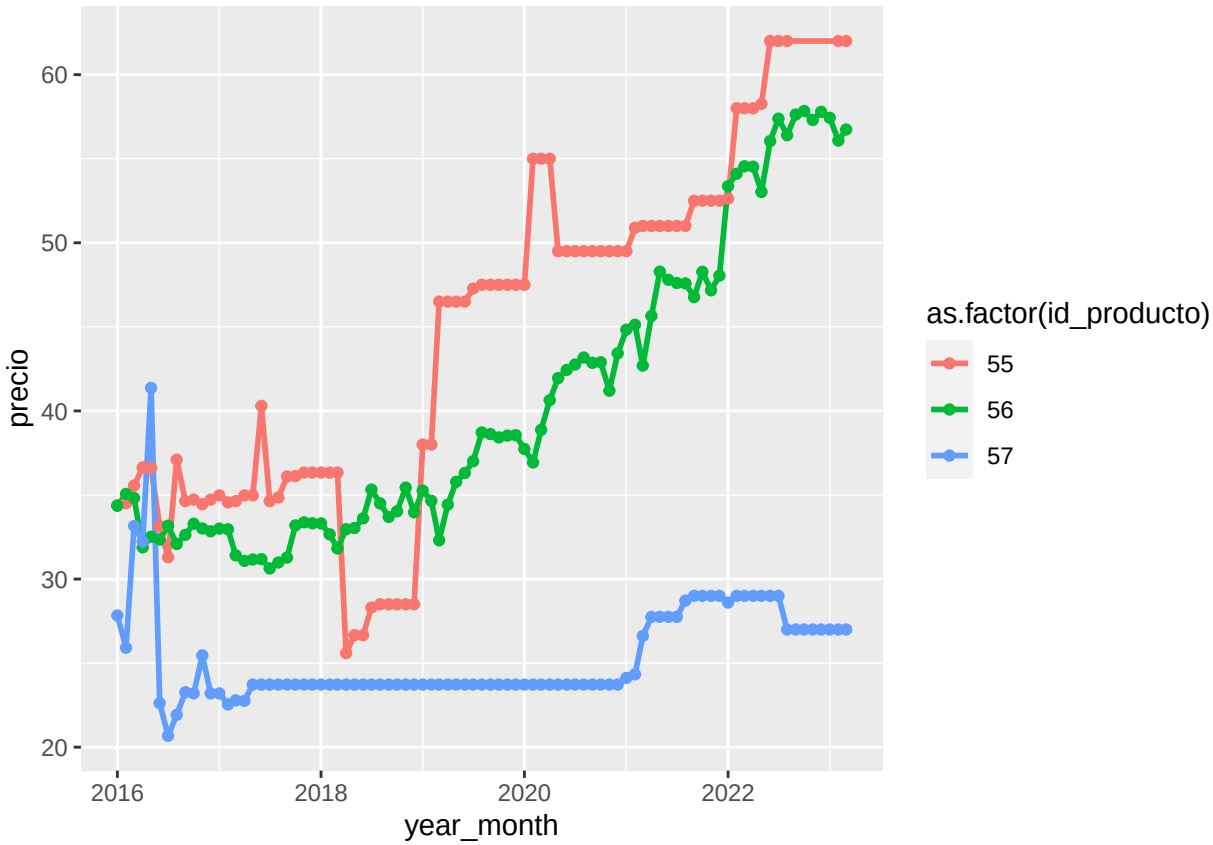
Se cuenta con datos a partir del año 2016 y hasta marzo del 2023. Estos contienen 363 productos dentro 766 establecimientos. De esos, solamente se usarán los 18 productos dentro de la CBA más 2 productos que se agregaron por decisión personal (para cada producto existen varias marcas).
Los productos son:

- Aceite de girasol 900 cc
- Aguja vacuna 1 kg (con y sin hueso)
- Arroz blanco 1 kg
- Arvejas en conserva 300 g
- Azúcar blanco 1 kg
- Carne picada vacuna 1 kg ***
- Cocoa 500 g
- Dulce de leche 1 kg
- Fideos secos al huevo 500 g ***
- Galletitas al agua 140 g ***
- Harina trigo común 0000 1 kg
- Huevos colorados 1/2 docena
- Manteca 200 g
- Pan flauta 215 g
- Papel higiénico hoja simple 4 rollos 30 mts
- Pollo entero fresco con menudos 1 kg
- Pulpa de tomate 1 L
- Sal fina yodada fluorada 500 g
- Yerba mate común 1 kg ***
- Café (agregado)
- Fideos secos de sémola 500 g (agregado)

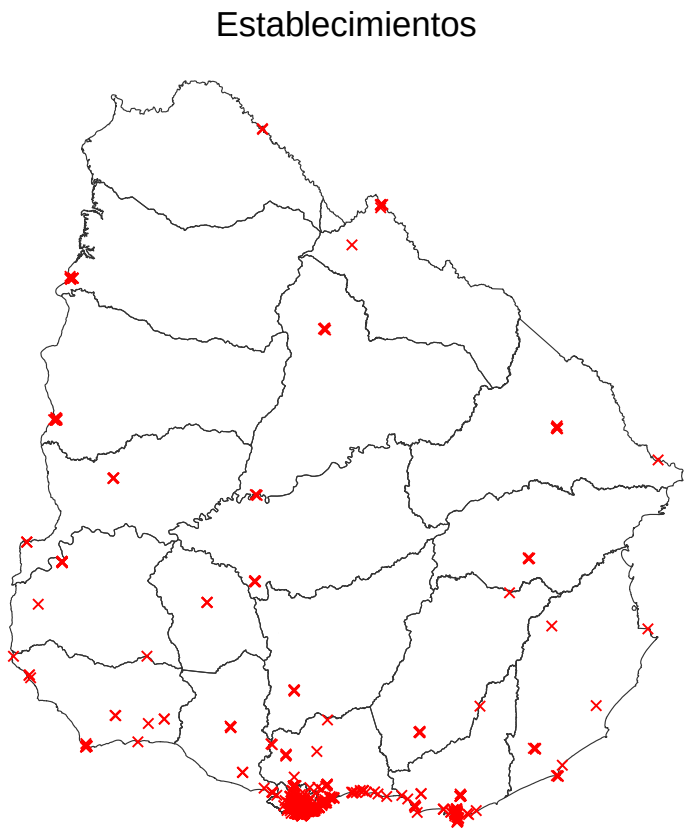
A partir de los datasets, se construyó uno con el precio promedio mensual para cada producto (desagregándolo según las marcas y establecimientos), que incluyera parte de la información presente en Establecimientos (el nombre de la sucursal, cadena, coordenadas, barrio y departamento) y Productos (nombre y marca). Para eso se utilizaron las *keys* respectivas, id.establecimientos e id_productos.

Análisis exploratorio

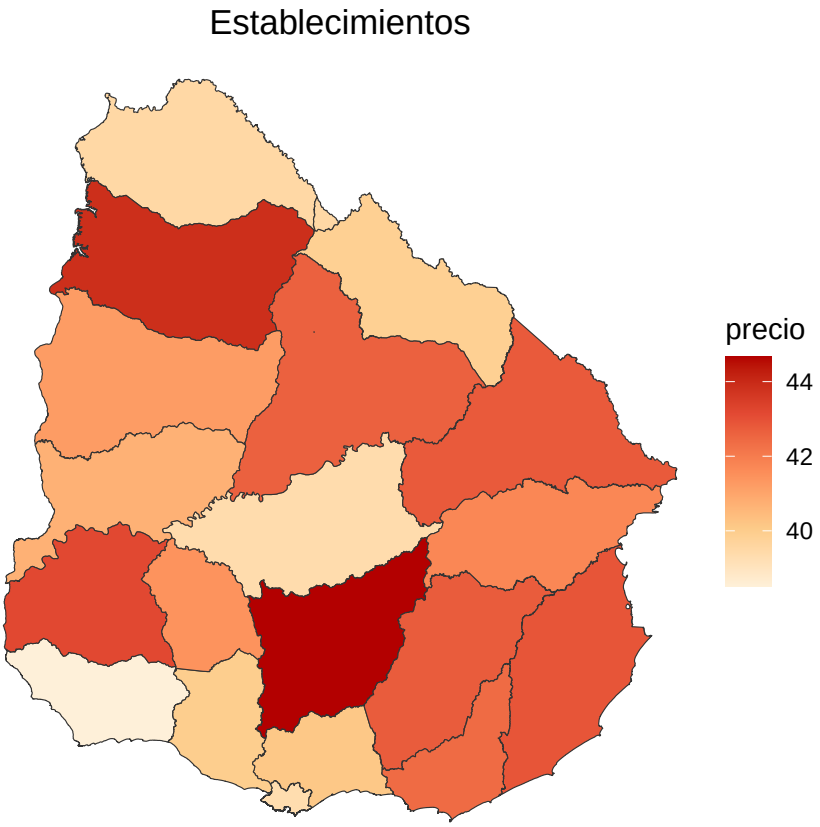
Evolución de precios de las distintas marcas de fideos



Mapa de los establecimientos



Mapa del precio promedio de los fideos según departamento



Modelo estadístico

En esta sección se describe el proceso de construcción de un modelo del precio de los productos en función del tiempo, que permita hacer predicciones sobre su variación en el futuro.

Para la confección del modelo, el producto seleccionado fue el paquete de fideos. Entre los productos trabajados, existen dos variedades de fideos: al huevo y semolados. A su vez, para cada variedad hay tres marcas.

En primer lugar se filtraron los datos para trabajar solamente con los precios del producto elegido. Para cada período (mes) se cuenta con precios por variedad, marca y por establecimientos. A su vez, todas las variedades y marcas registran el precio de paquetes de fideos del mismo tamaño (500 gramos). Por lo tanto, tuvo sentido agregar la información según varias, agrupaciones, calculando distintas "variables de respuesta":

1. Promedio general del precio de los fideos
2. Promedio del precio de los fideos para cada marca
3. Promedio del precio de los fideos según la ubicación del establecimiento que los distribuye (separando según Montevideo y el Interior)

En cuanto a las herramientas utilizadas, se exploraron varias alternativas hasta llegar a una satisfactoria. En primer lugar se escogieron los paquetes **caret** y **xgboost** para construir el modelo; y el paquete **forecast** para visualizar sus predicciones. El modelo se construyó en base a las instrucciones brindadas de este artículo. El problema de esta opción fue que el modelo creado no reconocía que la variable provista era el tiempo. Como indica la guía, la variable de tipo *Date* utilizada como explicativa se separa en dos: una indica el año

y la otra el mes de la fecha correspondiente. Sin embargo, para el modelo esas dos variables son simplemente numéricas, y lo que es peor, son tratadas como independientes entre sí.

Este problema se reflejó en predicciones que mostraban un patrón de variaciones que se repetía año a año, pero a pesar de la clara tendencia de aumento de los precios a lo largo del tiempo, los predichos se mantenían en un mismo rango. A continuación se muestra el resultado de las predicciones utilizando esta herramienta para el caso del promedio general de precios.

A continuación se muestran las predicciones obtenidas en cada caso

1. Modelo general del precio de los fideos

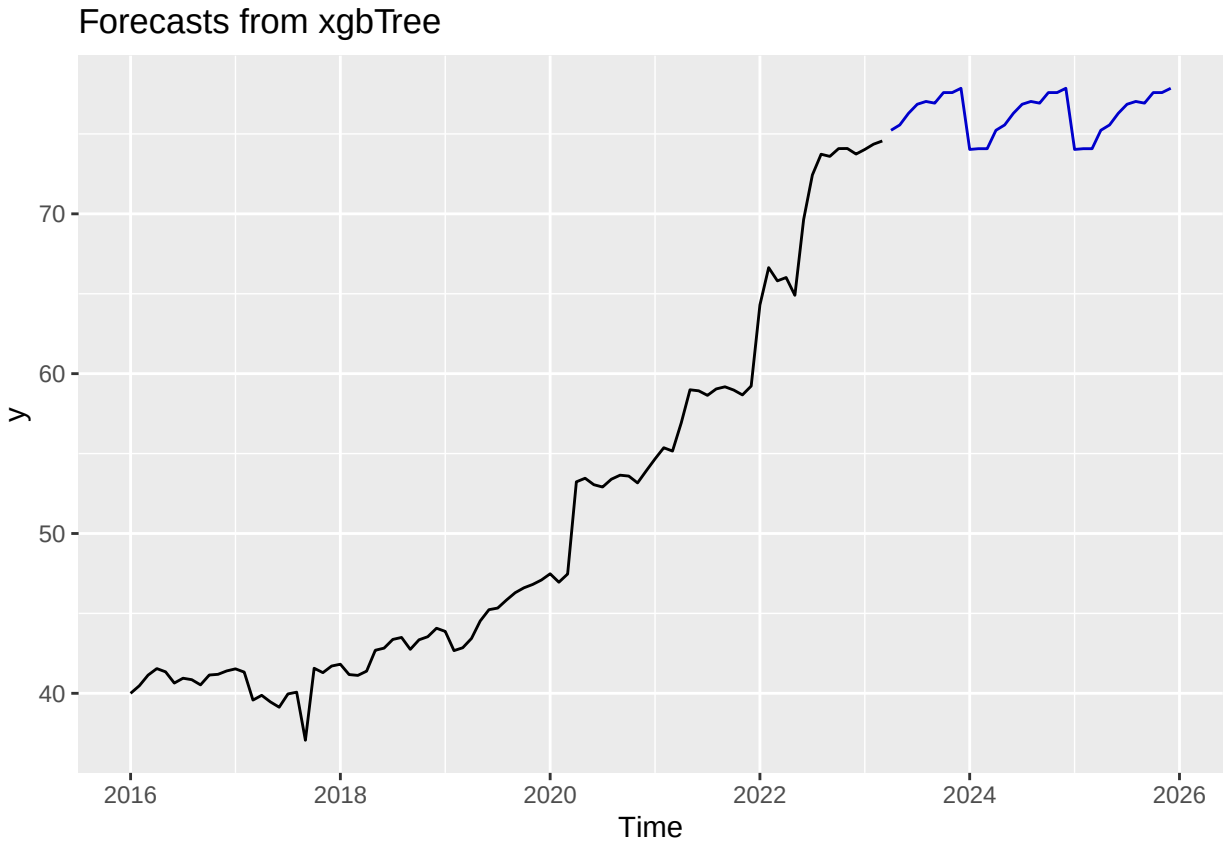


Figure 1: Prediccion del precio promedio de los fideos. El tramo en azul corresponde al período predicho.

En segundo lugar se optó por los paquetes **modeltime** y **timetk** para construir las series de tiempo. También se utilizó **tidymodels**, que brinda la posibilidad de utilizar la gramática de tidyverse en la confección de modelos. Estos paquetes difieren de los utilizados anteriormente en el sentido de que están específicamente diseñados para modelar con series de tiempo. A modo de ejemplo, en este caso la variable explicativa puede utilizarse en su clase natural (*Date*), en vez de transformarla a una o más numéricas.

Se realizaron 3 regresiones: ARIMA, GLMNET y Prophet. Para ello se seleccionaron los conjuntos de entrenamiento (*training*) y prueba (*test*). Para el primero se tomaron los datos desde el primer mes disponible hasta diciembre de 2022. Para el segundo se tomaron los primeros 3 meses de 2023 (los últimos períodos para los que hay datos). Los demás parámetros de cada regresión de dejaron con sus valores por defecto. Una vez creados los modelos, se utilizó la función modeltime_table de **modeltime** para integrarlos en una tabla y así llevar a cabo un análisis de los tres en conjunto.

Pasando a la etapa de evaluación, se analizaron las diferencias entre el conjunto de prueba y las predicciones halladas. Luego se reajustó el modelo considerando la totalidad de los datos y se calcularon las predicciones para los 9 meses restantes de 2023. Además de gráficos interactivos con los valores predichos para cada regresión, el paquete **modeltime** cuenta con la función *modeltime_accuracy*, que muestra una serie de medidas para evaluar sus desempeños.

A continuación se presentan imágenes con los resultados obtenidos para cada regresión en el caso del precio promedio general, y la tabla con las medidas:

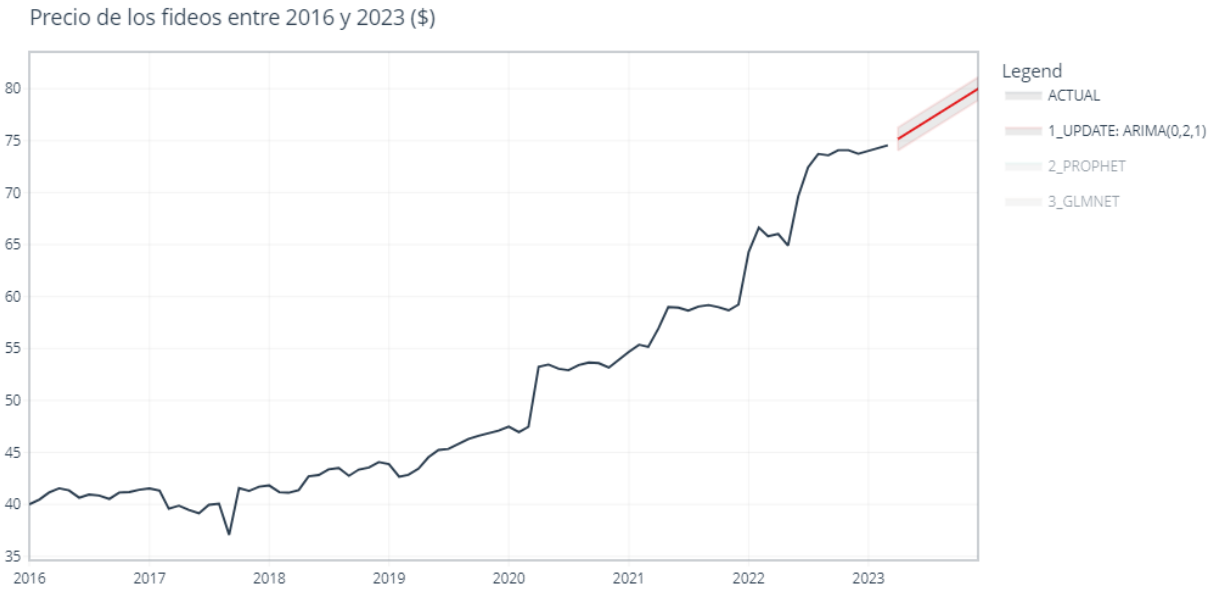


Figure 2: arima

	modelo	mae	mape	mase	smape	rmse
1	ARIMA(0,0,0) WITH NON-ZERO MEAN	0.47	0.63	1.77	0.63	0.52
2	PROPHET	1.28	1.72	4.83	1.74	1.37
3	GLMNET	0.83	1.11	3.11	1.12	0.85

Como se puede ver, los tres modelos predicen que el precio de los fideos va a aumentar en los próximos 9 meses.

En cuanto a la tabla, las medidas corresponden a distintas formas de agregar el *error de predicción*, que se define como la diferencia entre los datos del conjunto de prueba y las predicciones del modelo. Según el libro **Forecasting: Principles and Practice** se pueden clasificar en tres tipos: - Dependientes de la escala (de la variable respuesta): mae y rmse - Errores porcentuales: mape y smape - No dependientes de la escala: mase

El primer tipo sirve exclusivamente para comparar modelos en los que la variable respuesta esté medida en una unidad común, ya que el error se mide en las unidades de la misma. Los otros dos sirven para comparar el desempeño de modelos más diversos.

En este caso, como los tres modelos están construidos con el mismo conjunto de datos, se pueden considerar todos los estadísticos. En todos los casos el error más pequeño lo tiene la regresión de ARIMA, por lo que se concluye que es la mejor alternativa para esta serie de tiempo.

En cuanto a su intervalo de confianza, a un 90% tiene un rango de aproximadamente 2 pesos, que teniendo en cuenta el problema que se está estudiando, no presenta mayores inconvenientes.

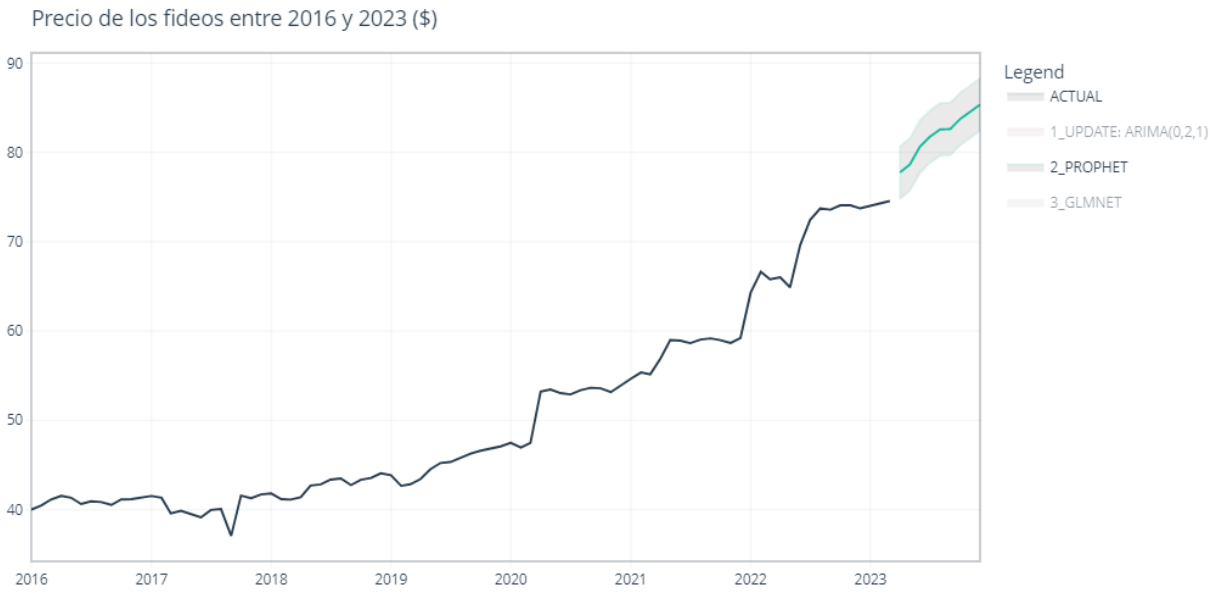


Figure 3: prophet

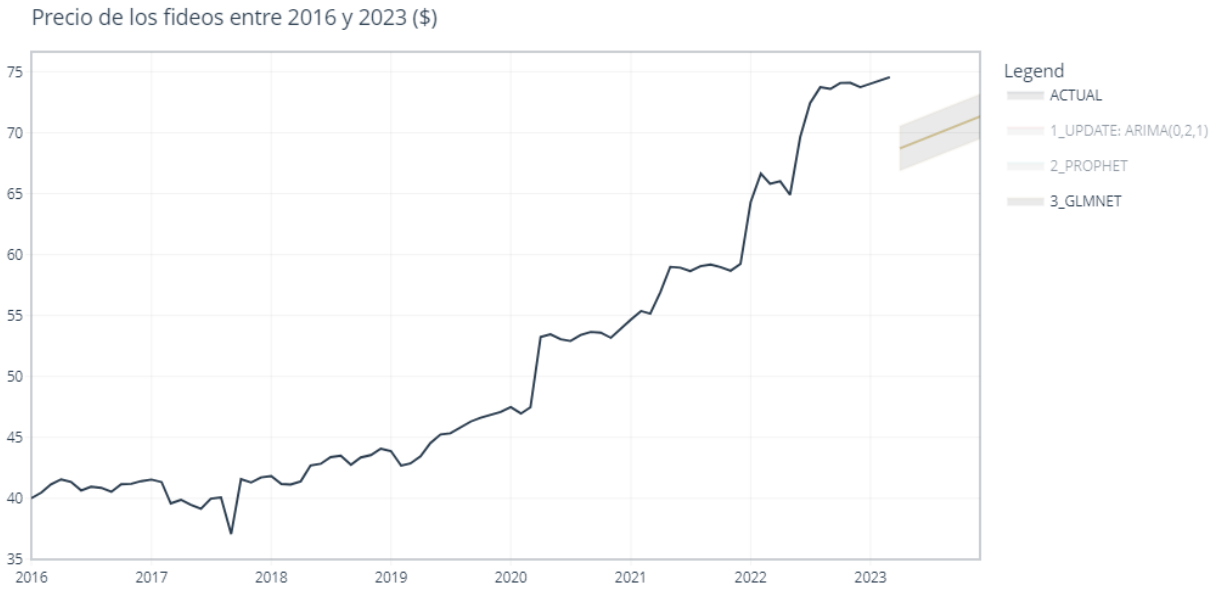


Figure 4: glmnet

A modo de conclusión, el modelo predice que al final del año 2023 el precio de los fideos habrá experimentado un aumento de 7.80% con respecto al año anterior. Este aumento está alineado con la inflación esperada para el año, que es de un 8.00% según la Encuesta de Expectativas Empresariales del INE realizada en junio del presente año.

Aplicación Shiny

Referencias

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2022). *Shiny: Web application framework for r.* https://CRAN.R-project.org/package=shiny

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & Yuan, J. (2023). *Xgboost: Extreme gradient boosting.* https://CRAN.R-project.org/package=xgboost

Cheng, J., Karambelkar, B., & Xie, Y. (2023). *Leaflet: Create interactive web maps with the JavaScript 'leaflet' library.* https://CRAN.R-project.org/package=leaflet

Dancho, M. (2023). *Modeltime: The tidymodels extension for time series modeling.* https://CRAN.R-project.org/package=modeltime

Dancho, M., & Vaughan, D. (2023). *Timetk: A tool kit for working with time series.* https://CRAN.R-project.org/package=timetk

Hyndman, R., & Athanasopoulos, G. (2018). 3.4 evaluating forecast accuracy. In *Forecasting principles and practice.* OTEXTS.

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org

Kuhn, & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E., & Bivand, R. (2023). *Spatial Data Science: With applications in R* (p. 352). Chapman and Hall/CRC. https://r-spatial.org/book/

R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686