Reporte de selección y parametrización de modelos

Cristhian Amaya / Andrés Beltrán / Rodrigo Méndez

1. Tratamiento de datos:

1.1 Datos Frubana:

Frubana nos compartió 5 bases de datos iniciales.

Base de datos	Formato	Descripción
Forecast	1 archivo pkl	Tabla que contiene la información sobre el pronóstico que se realiza para cada uno de los productos
BAQ_compras	1 archivo csv	Esta base contiene las compras que llegaron a la Bodega de Barranquilla entre el 26 de julio de 2022 hasta el 29 agosto de 2023
Productos_BAQ	1 archivo csv	Muestra las características de los productos que se compran en la bodega de barranquilla
Ventas	12 archivos pkl	Las bases contienen información sobre las ventas de Frubana a sus clientes.
Waste percentage by age	1 archivo csv	La base contiene información sobre el porcentaje de pérdida del producto con respecto al número de días que han pasado desde que se obtuvo el producto.

Y con estos archivos en físico, se procedió a crear un notebook en Jupyter donde se creó un pipeline que reviso:

- La calidad de los datos
- Las dimensiones de cada tabla
- Los fragmentos de tiempo de cada tabla

Nota: Estas evaluaciones iniciales de los datos se pueden encontrar en (https://github.com/cramayag/Pricing_Frubana/blob/main/Datos%20Frubana/Analisis%20exploratorio.ipynb).

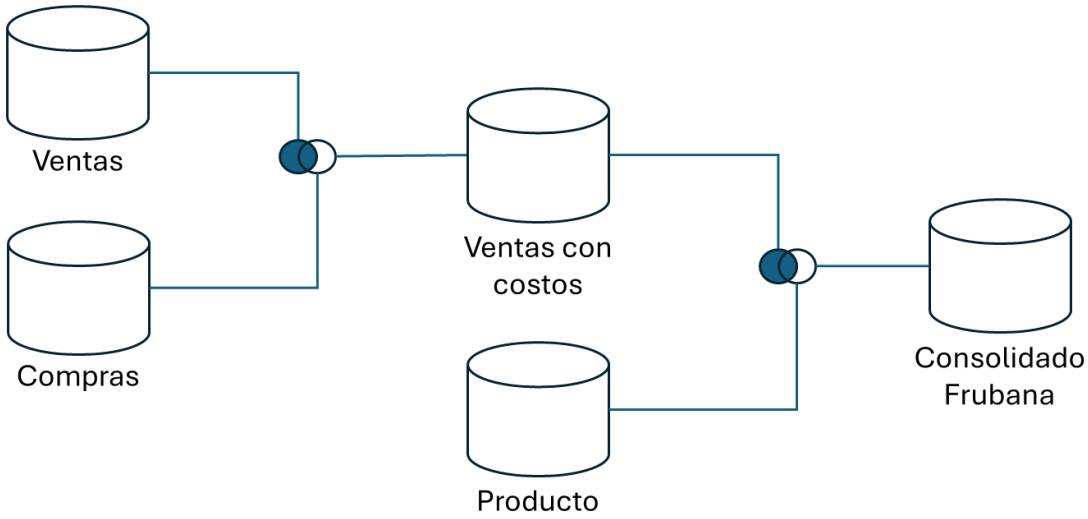
Después de evaluar los datos, se debe pasar a la consolidación de una única base de datos que permita empezar a trabajar con los modelos a desarrollar para responder la pregunta del ejercicio. Como se busca evaluar la contribución de cada producto, y considerando que la información de costos no se compartió de manera implícita, trabajaremos con las bases de datos de compras y ventas bajo los siguientes supuestos:

- El fragmento de tiempo de ambas bases debe ser el mismo. Esto implico acotar ambas bases de datos, puesto que ventas contiene registros desde 1 de Julio del 2022 al 1 de Julio del 2023, y la base de datos de compras contiene datos desde el 26 de Julio del 2022 hasta el 29 de agosto del 2023.

- Vamos a considerar la base de datos de compras como la base de costos que unificaremos con la base de ventas. Como la base de datos de compras tiene compras por fecha, vamos a trabajar con el supuesto de que un producto está en promedio 4 días en inventario. Lo que quiere decir que el precio de compra de un producto i de hace 4 días será el costo de venta del mismo producto i en la base de ventas.
- Hay productos en base de datos de ventas que no lo hacen, por lo que, como no pueden ser costeados, no se considerarán en la base de datos consolidada.
- Los costos operacionales relacionados con el inventario, mano de obra, etc, se estiman como el 14% de compras totales del día. De esta manera se distribuirá en los costos finales en cada producto.

Para más detalles de las transformaciones realizadas (https://github.com/cramayag/Pricing_Frubana/blob/main/Datos%20Frubana/Analisis%20exploratorio.ipynb).

Después de realizar la unión de ambas bases de datos, se realizó un último join con la base de datos de productos, que nos ofrecerá más valores para la tabla consolidada. De esta manera, se llega a la base de datos consolidada con datos de Frubana.



La base consolidada de Frubana tiene las siguientes características:

- Forma: 325,445 X 29
- Tamaño: 9’437,905
- El producto que más frecuentemente aparece en registros es la “Cebolla Cabezona Blanca Sin Pelar Mixta Desde 1Kg”
- El precio medio en la base de datos es de $4,805, con un min de $90, una mediana $3,297 y un máximo de $100,000.
- En promedio se dieron descuentos registrados en $235.
- La unidad de producto más usada en venta es el Kg.

- La categoría más frecuente es verduras (69% de registros), seguido por las frutas (18.4%) y Tubérculos (10.3%)

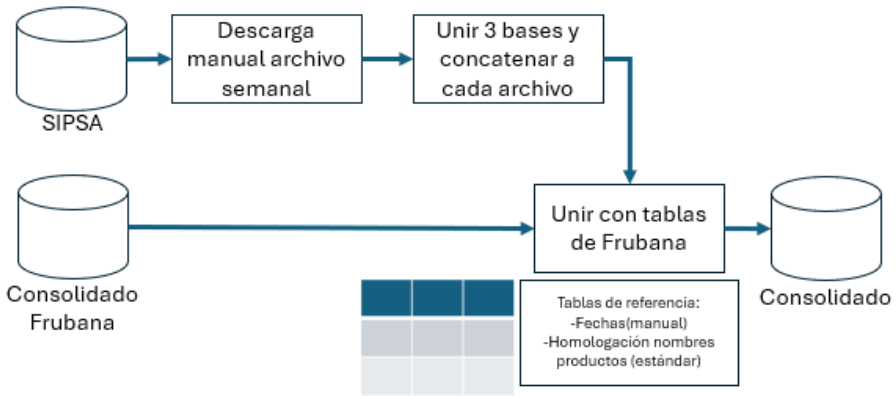
1.2 Datos externos:

El Departamento Administrativo Nacional de Estadística (DANE), a través del Sistema de Información de Precios y Abastecimiento del Sector Agropecuario (SIPSA), proporciona los precios de verduras y frutas de diversas ciudades del país. Mediante el reporte semanal, es posible descargar manualmente el precio medio de la información, y mediante código, se puede realizar un barrido de la carpeta para concatenar la información útil para el proyecto. Repositorio (https://github.com/cramayag/Pricing_Frubana/blob/main/Datos%20Externos/code/pipeline%20externo-checkpoint.ipynb)

El informe del SIPSA se divide en diversas hojas, incluyendo Verduras y hortalizas, Frutas frescas, Tubérculos, raíces y plátanos, Granos y cereales, Huevos y lácteos, Carnes, Pescados, Productos procesados, Arroz y subproductos en molino, y Abastecimiento semanal por grupo de alimentos.

Para el funcionamiento del código, se utilizan las tres primeras hojas, donde se concatenan los datos de 113 archivos de Excel. Estos archivos contienen las siguientes columnas: 'Producto', 'Mercado mayorista', 'Precio mínimo', 'Precio máximo', y 'Precio medio' (ver archivo unificar_semanas_sipsa.py). Además, se trabajan con dos archivos de referencia, que son realizados a partir comparando la información manual: 'fechas.xlsx', que contiene el nombre y rango de fecha de los precios, y 'homologación_sipsa_frubana_1', que realiza la homologación del nombre SIPSA al nombre del sistema de información de Frubana.
Repositorio(https://github.com/cramayag/Pricing_Frubana/blob/main/Datos%20Externos/code/unificar_semanas_sipsa.py)

Con esta información se realiza la consolidación de la base de datos de Frubana realizando el cruce con los precios de SIPSA de la ciudad de Barranquilla teniendo en cuenta las tablas referenciales.



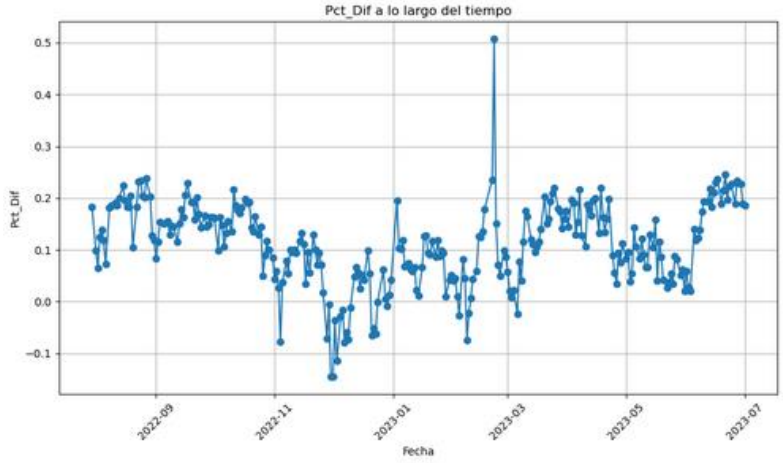
A partir de esta conexión de datos se logró examinar las correlaciones dadas por los precios de cada producto versus los demás y las correlaciones dadas entre las variables finales.

2. Verificación de modelos utilizados:

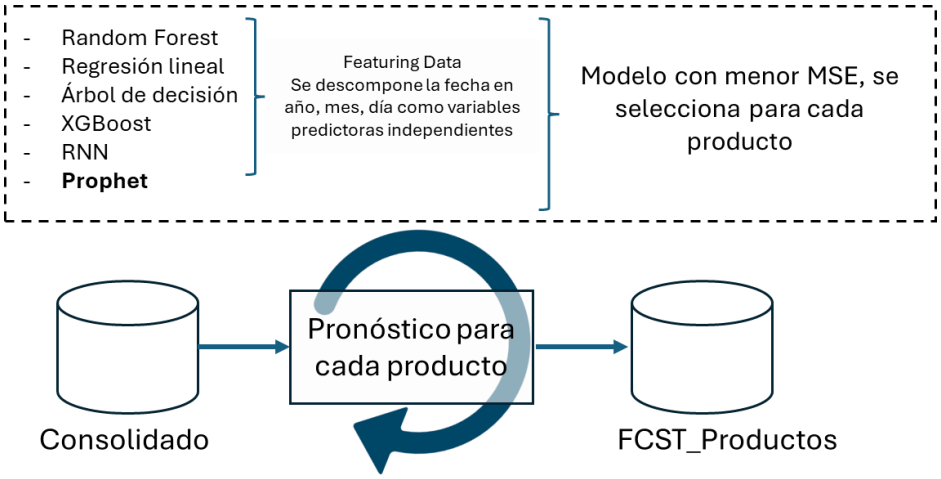
2.1 Modelo de apoyo pronóstico precios de Mercado

A partir de la unificación de la información del SIPSA y la empresa se revisa el porcentaje de diferencia del precio de los productos de la empresa frente al precio medio del mercado para cada producto diario. Al realizar un análisis de la información entre ellos ver el promedio de los productos por día se presenta el siguiente gráfico:

Repositorio (https://github.com/cramayag/Pricing_Frubana/blob/main/Datos%20Externos/code/pipeline%20externo.ipynb)



Esto evidencia que, en promedio, Frubana tiene precios por encima de los precios de mercado. Para realizar una comparativa en los próximos días, se realiza un pronóstico basado en la información del SIPSA. Este pronóstico compara, producto por producto, la métrica del Error Cuadrático Medio (MSE) de 6 modelos de pronóstico. Se ha observado que cada producto tiene un comportamiento diferente, y el mejor modelo que describe el precio de un producto no describe el precio de otro por ende se evaluó inicialmente estos modelos.



Para el uso de 5 modelos iniciales, se manejan como variables independientes el día, mes y año de la información. A partir de estos datos, se descomponen y se ingresan como variables independientes en los modelos. En el caso de Prophet, se utiliza la fecha como parámetro "ds", ya que este modelo está diseñado para manejar series de tiempo.

Para el Ñame, por ejemplo, se evaluaron los siguientes modelos, donde el modelo con menor MAE y MSE es el modelo de Gradient Boosting. Al comparara los distintos modelos se encontró que el Gadient Boosting era seleccionado en los diferentes productos.

```
Arbol MAE  840.9754028837998
Arbol  MSE  1286209.335029686
Regresión Lineal MAE  974.9092359507464
Regresión Lineal MSE  1422393.6766397627
RF MAE  734.5670802638778
RF MSE  940578.0883880438
GB MAE  676.3375101325158
GB MSE  808911.198876549
RNN MAE  986.1875339180723
RNN MSE  1431283.0631625832
```

```
Clave
G Boosting     180
Arbol            1
Regresion_l      1
```

Repositorio(https://github.com/cramayag/Pricing_Frubana/blob/main/Datos%20Externos/code/pipeline%20externo.ipynb)

Por ende, se perfeccionaron los distritos modelos teniendo como resultado un mejor desempeño un Random Forest con una búsqueda aleatoria de hiperparámetros para cada uno de los productos, evaluando múltiples combinaciones de hiperparámetros y seleccionando el modelo que obtiene el mejor rendimiento en términos de la métrica especificada. Por ejemplo, en el caso de "Acelga", Para "Ahuyama", el MAE del Random Forest (247.24) fue significativamente menor que el del árbol de decisión (313.41), lo que indica que el Random Forest tuvo un rendimiento mucho mejor en términos de error absoluto medio. Los hiperparámetros que se incluyeron para mejorar el modelo son el número de estimadores (árboles) en el bosque (n_estimators), la profundidad máxima de cada árbol (max_depth), el número mínimo de muestras requeridas para dividir un nodo interno

(min_samples_split), el número mínimo de muestras requeridas para ser un nodo hoja (min_samples_leaf), y la cantidad máxima de características a considerar en cada división (max_features). Estas configuraciones permiten explorar diversas combinaciones para optimizar el rendimiento del Random Forest en la tarea de modelado.

2.2 Modelo principal para pricing

Para el modelo de pricing, tomamos los 5 productos con mayor número de ventas para empezar a modelar el comportamiento de sus respectivas demandas (variable a predecir) y los cambios de precios (variable predictora).

En esta etapa de modelado, se probarán diferentes técnicas de regresión para determinar la relación de cantidad vendida vs precios o cantidades de otros productos, la búsqueda permitirá determinar una razón de cambio entre un precio asignado y su venta. Una parte fundamental del desarrollo de modelos es la selección de las variables importantes para el modelo por lo cual se desarrolla una metodología lógica para identificar estas variables.

El proceso de selección de variables se hace basado en una metodología que tienen cuenta los siguientes pasos:

1. Identificación de productos con **mayor correlación** respecto a la variable objetivo a modelar (cantidad vendida de producto): una vez generadas las matrices de correlaciones de cantidad y precios, se seleccionan los productos que mayor correlación tiene respecto al producto a modelar, con esto, se seleccionan 10 productos más correlacionados con el precio y 10 productos con mayor correlación con la cantidad. Si se repiten los productos, se elimina el de menor correlación entre ambos.
2. Con esta selección de variables previa, se construye la tabla base y se calculan las correlaciones entre las variables con el uso de un correlograma. Aquellas variables cuya **correlación sea nula**, serán filtradas.
3. Utilizando el estadístico VIF (Factor de Inflación de Varianza) calculado para cada variable de la tabla base, quedarán dentro de la selección aquellas variables que su **VIF sean inferior a 10**, evitando tener problemas de colinealidad en la construcción de modelos.

Es importante mencionar que la base original tiene información aproximadamente 280 días, al no tener un gran volumen de información por producto para ser analizada, se decidió no hacer una partición de entrenamiento y comprobación sino crear los modelo con Validación cruzada, y en algunos casos, usando Kfold-validation. Con lo anterior se entrenarán modelos que permiten usar todos sus datos.

Definidas las variables que serán usadas en la construcción de los modelos, definiendo el producto a predecir el precio (Ñame), se probaron diferentes modelos como:

Modelo 1 Regresión lineal base con Kfold-Validation (LR):

Este modelo base, introduce todas las variables seleccionadas anteriormente, en la construcción del modelo se generan 10 folds para ser usado en la técnica de Kfold-validation, y de los datos origen se hacen muestreos y se corre k modelos promediando el resultado final de las métricas.

```
Average MSE across all folds: 32799.88324915704
Average R2 across all folds: 0.07725369909484145
```

Modelo 2 Red neuronal con Kfold-Validation (NN):

Este modelo se crea con tres capas ocultas con 20, 20, 20 neuronas respectivamente, Kernel de regularización L2 (Lasso) y una capa de salida con una función de activación linear. Los resultados de este modelo no fueron favorables, pero se muestran los obtenidos.

```
Average MSE across all folds: 43768.391146320166
Average R2 across all folds: -0.2817465514939222
```

Modelo 3 Modelo GAMS (GAMS):

Se entrena un modelo GAM utilizando Kfolds-validación, sin embargo, los resultados obtenidos por este modelo no mejoraban respecto a los modelos revisados anteriormente.

```
Average MSE across all folds: 33253.046756873984
Average R2 across all folds: 0.07410865945825997
```

Modelo 4 Regresión lineal usando validación cruzada (LR2):

Para este entrenamiento de modelo se utiliza validación cruzada y Kfold usando toda la información de los datos de entrada, el modelo entrenado genera mejores resultados que los modelos anteriores debido al proceso de validación que nos permite el uso de todos los datos para entrenamiento del modelo.

```
Average MSE across all folds: 26471.84372567028
Average R2 across all folds: 0.3085780009762954
```

Modelo 5 Regresión lineal con Kfold y Stats models (LR Stats):

Este modelo se entrena con el paquete Stats models que utiliza para los modelos de regresión lineal OLS (mínimo cuadrados ordinarios), la generación de resultados de modelado usando esta técnica permite tener resultado interesantes para la naturaleza de los datos.

```
MSE promedio en validación cruzada: 29283.834613128587
R2 promedio en validación cruzada: 0.34164625364678686
```

Modelo 6 Regresión lineal usando Pasos hacia delante (SFS) y Kfold (LRSFS):

Para el entrenamiento de este modelo se tiene en cuenta el método de pasos hacia delante, donde el modelo selecciona de las variables disponibles aquellas variables significativas, adicional se tiene en

cuenta Kfold validación con el fin de entrenar y validar el modelo. Los resultados obtenidos con esta técnica se reducen a los obtenidos anteriormente.

```
MSE promedio en validación cruzada: 34848.97826146886
R2 promedio en validación cruzada: 0.018331678229532012
```

Modelo 7 Calibración modelo GAM (GAMS_2):

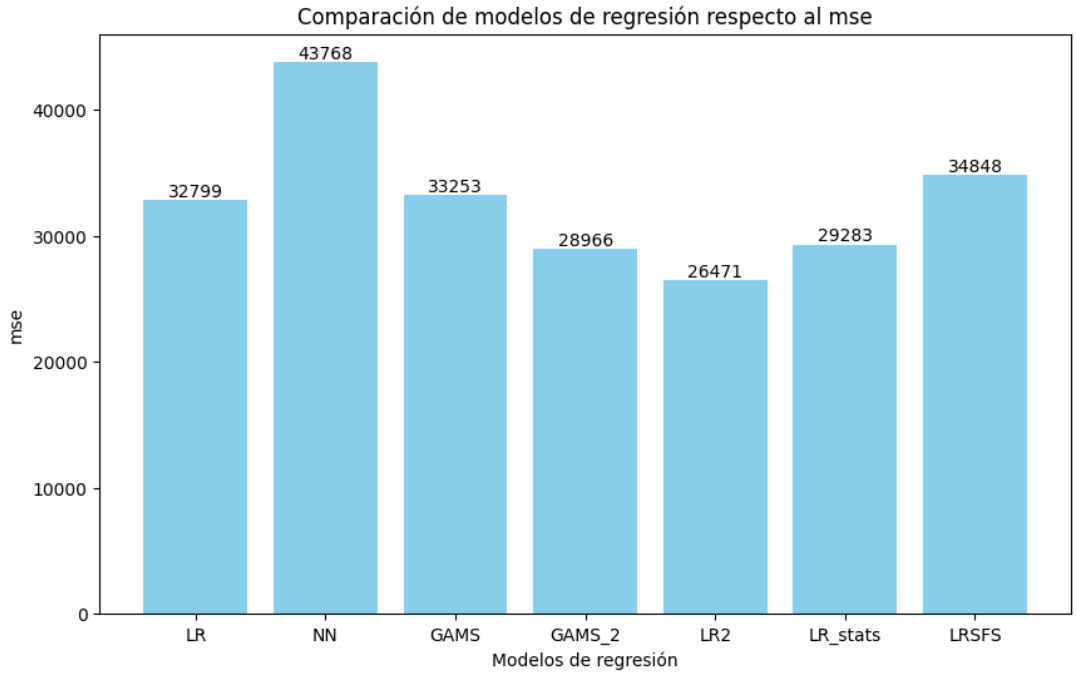
Para este entrenamiento, se desarrolla el modelo gam con mejoras de calibración como la definición de cada uno de los términos, utilizando spines cúbicos y suaves para modelar las relaciones no lineales entre las variables de entrada y la variable de salida. Los resultados que se obtienen de la parametrización fue la siguiente.

```
Average MSE across all folds: 28966.751478715178
Average R2 across all folds: 0.18279491003994525
```

El detalle de la construcción de los modelos se encuentra en el repositorio de gitHub (https://github.com/cramayag/Pricing_Frubana/blob/main/Modelos%20Pricing/Model/Modelo_de_pricing.ipynb)

3. Análisis resultados e implementación:

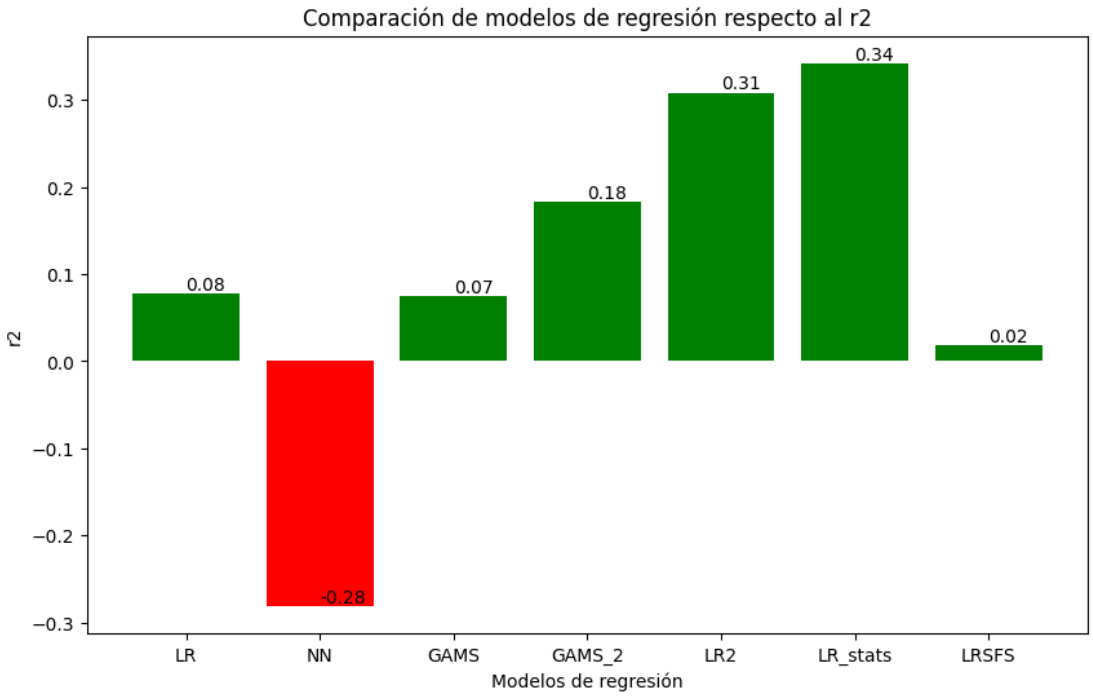
Las métricas que se tendrán en cuenta para seleccionar los modelos que ayudan a encontrar la relación entre precio y demanda de producto, son el **MSE** y **R^2,** con estas métricas, se seleccionaran los modelos con menor MSE y mayor R^2, estas métricas se presentan en los gráficos a continuación:



El gráfico muestra el comportamiento de los modelos probados, a cada modelo se les calculó el MSE y con esta métrica, se comparan para determinar aquellos que tuvieron mejor desempeño. El modelo

con mejor comportamiento es el de regresión lineal con validación cruzada el cual arroja un mse=26.471 y el de menor ajuste es la red neuronal con un mse=43.786.

Adicional se muestran los resultados de los modelos respecto a R^2 para cada uno de los modelos



Los modelos con mejor desempeño respecto a R^2 son: modelo de regresión con validación cruzada R^2 = 0.31 y modelo de regresión lineal que utiliza el método de mínimos cuadrados ordinarios con un R^2 = 0.34 los cuales serán modelo para desplegar con los demás productos seleccionados. Las redes neuronales son las que peor desempeño tiene con un R^2 negativo lo cual indica que el resultado es peor al promedio de los datos.

Con las métricas planteadas anteriormente, el criterio se basa en que en ambas métricas tengan un buen desempeño para ser candidatas al desarrollo para los demás productos.

Análisis de resultados del modelo con mejor métricas de desempeño:

Una vez analizados los resultados de los modelos con las métricas de desempeño, se profundizan en los resultados obtenidos por el **Modelo 5** Regresión lineal con Kfold y Stats models (LR Stats), el cual arroja mejores resultados ajustándose mejor a los datos.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.342
Model:	OLS	Adj. R-squared:	0.296
Method:	Least Squares	F-statistic:	7.411
Date:	Mon, 06 May 2024	Prob (F-statistic):	3.95e-13
Time:	03:58:34	Log-Likelihood:	-1496.2
No. Observations:	230	AIC:	3024.
Df Residuals:	214	BIC:	3079.
Df Model:	15		
Covariance Type:	nonrobust		

Coeficiente de las variables

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.2040	0.053	-3.877	0.000	-0.308	-0.100
x2	-0.0122	0.052	-0.233	0.816	-0.115	0.091
x3	-0.0430	0.072	-0.595	0.552	-0.186	0.099
x4	-0.1113	0.049	-2.279	0.024	-0.208	-0.015
x5	-0.0179	0.008	-2.131	0.034	-0.034	-0.001
x6	0.0240	0.023	1.029	0.304	-0.022	0.070
x7	-0.0081	0.011	-0.761	0.447	-0.029	0.013
x8	0.0014	0.008	0.167	0.868	-0.015	0.018
x9	0.0190	0.010	1.826	0.069	-0.002	0.039
const	0.1266	0.081	1.564	0.119	-0.033	0.286
x10	0.0038	0.070	0.054	0.957	-0.134	0.141
x11	0.0318	0.058	0.543	0.588	-0.084	0.147
x12	0.0089	0.008	1.147	0.253	-0.006	0.024
x13	-0.0718	0.031	-2.351	0.020	-0.132	-0.012
x14	0.0106	0.015	0.683	0.495	-0.020	0.041
x15	0.1800	0.115	1.564	0.119	-0.047	0.407
x16	-0.0318	0.026	-1.199	0.232	-0.084	0.020

Omnibus:	8.635	Durbin-Watson:	2.117
Prob(Omnibus):	0.013	Jarque-Bera (JB):	10.357
Skew:	0.314	Prob(JB):	0.00564
Kurtosis:	3.829	Cond. No.	3.13e+16

Finalmente, los resultados obtenidos con este modelo para el producto de Ñame son:

```
Average MSE across all folds: 29283.834613128587
Average R2 across all folds: 0.34164625364678686
```

Con la identificación de estos modelos y las métricas de desempeño a usar, se procede a realizar el modelado de datos para los demás productos.

4. Siguientes pasos

A continuación, se listan los puntos a abordar en las siguientes semanas para lograr resultados de los objetivos del proyecto.

- Ejecución de modelos seleccionados para los productos más representativos de Frubana.
- Consolidación de modelos de pronósticos y modelos de pricing para optimizar la contribución esperada de cada producto a modelar.
- Modelo de optimización para productos para definir el precio que maximiza el beneficio.
- Construcción de un tablero para visualizar los resultados obtenidos del proyecto respecto a la asignación de precios.
- Manual de despliegue de la solución de manera local, desde la importación de datos hasta a la obtención de resultados de los modelos para su respectiva toma de decisiones.