Aplicación de Machine Learning a Predicción de Demanda

Gutiérrez Montecino, Denise
Nieva, Sofia
Rodríguez, Alfredo

22 de Octubre 2021

Hoja de Ruta

1. Introducción y Objetivos
2. Análisis Exploratorio y Curación de Datos
3. Machine Learning
4. Clusterización
5. Desafíos y Conclusiones

Introducción y Objetivos

Introducción

- ✔ Predicción de Demanda de productos congelados
- ✔ Datos de ventas de productos elaborados por una compañía de alimentos congelados en distintos países de la región entre 2018-2019

Objetivos

- ✔ Predecir cuál será la venta de los productos para los próximos meses en las diferentes zonas donde opera la compañía

Análisis Exploratorio y Curación de Datos

1. Introducción y Objetivos
2. Análisis Exploratorio y Curación de Datos

Descripción del Dataset

- ✔ Nuetros set de datos se compone de 5 datasets:
  1. Productos
  2. Categorías
  3. Puntos de Venta
  4. Ventas
  5. Países

- ✔ Se unificaron en un sólo dataset con un total de 5,5 millones aprox. de registros y 25 variables

- ✔ Dentro de las variables predictoras se seleccionó: variable de tiempo '**time**', variable de donde se originó el pedido '**Ubicacion**', variable que identifique el producto '**sku**' y los valores de venta que se captura con la variable '**totalkg**'

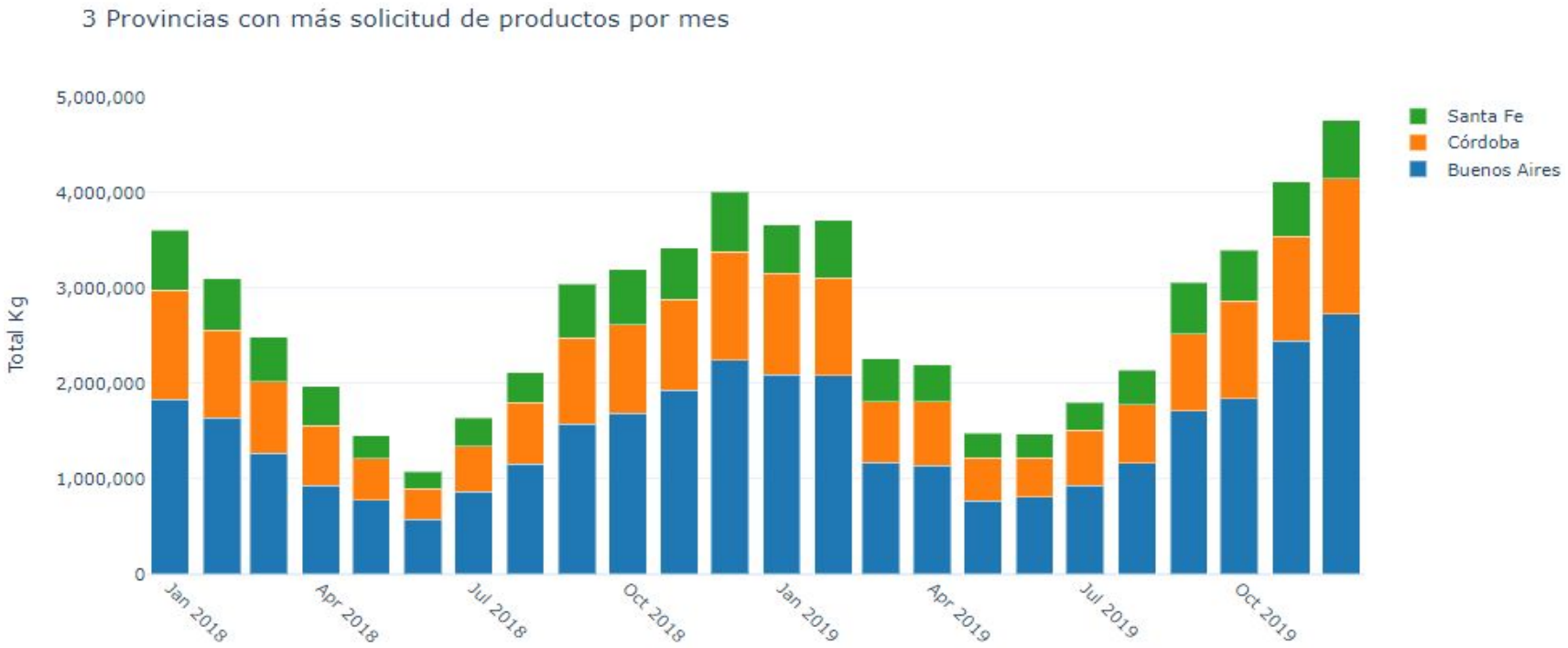
Análisis Exploratorio y Curación de Datos

- ✔ Generación de variable '**totalkg**' resultante del producto entre 'cantidad_pedida' y 'unidadkg'
- ✔ Se detectaron valores negativos y nulos, estos no son sistematicas y son no informativos a los fines del análisis por lo que se eliminan del dataset (equivalente al 5% aprox del dataset original)

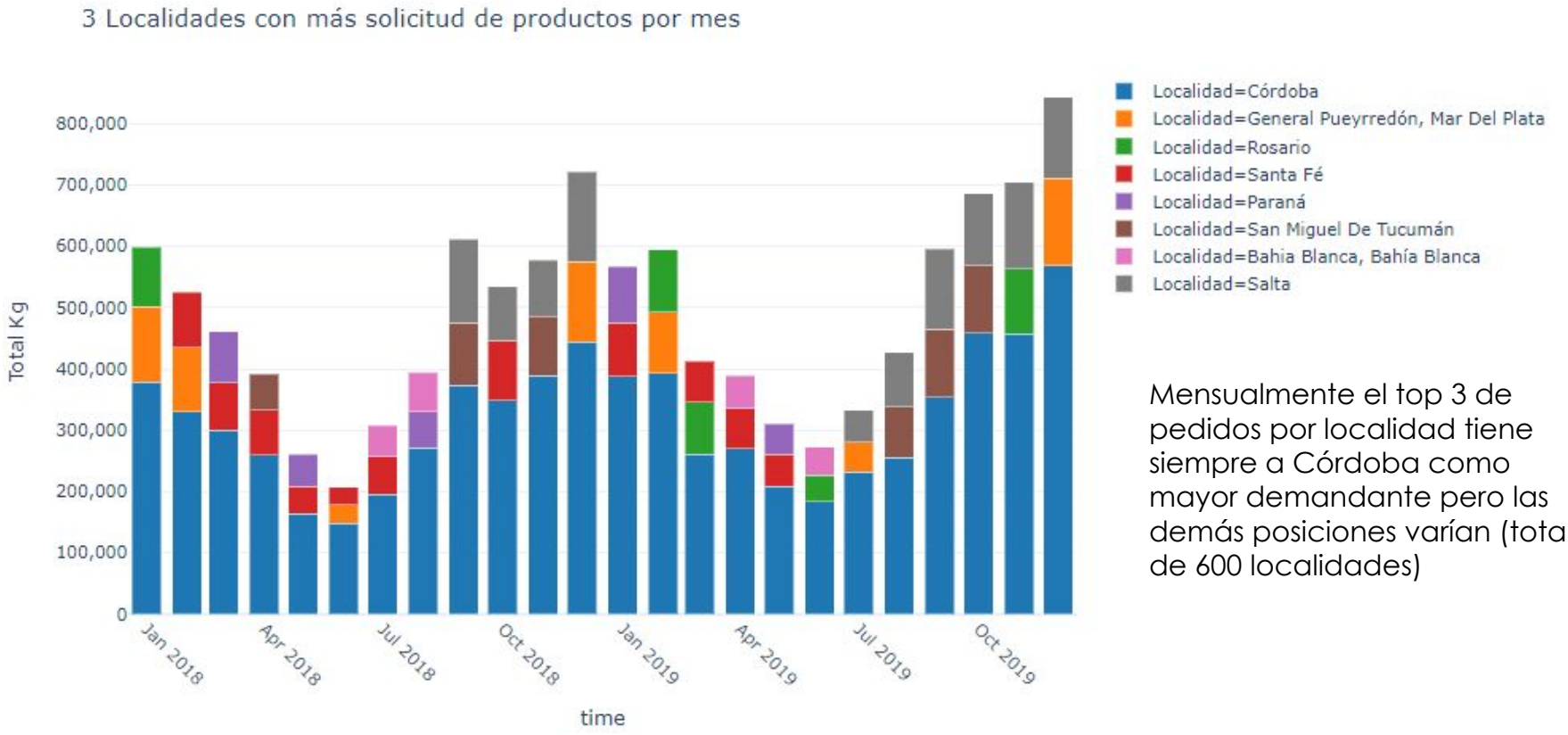
- ✔ Variable '**Categoria**': se detectaron categorías de productos no comestibles. Dichos productos presentan una alta densidad de datos anómalos en 'unidadkg'
- ✔ Dado que no aportan al objetivo del estudio se quitaron del dataset (8% aprox del dataset original)

- ✔ Se detectan valores faltantes para las variables de ubicación **'id_Provincia', 'id_Localidad', 'id_Pais', 'Localidad', 'Provincia'** y **'Pais'.**
- ✔ Las pérdidas son sistemáticas correspondientes a las ventas fuera de Argentina. Se eliminan del set de datos (11.78%)

Variables de interés: Provincias

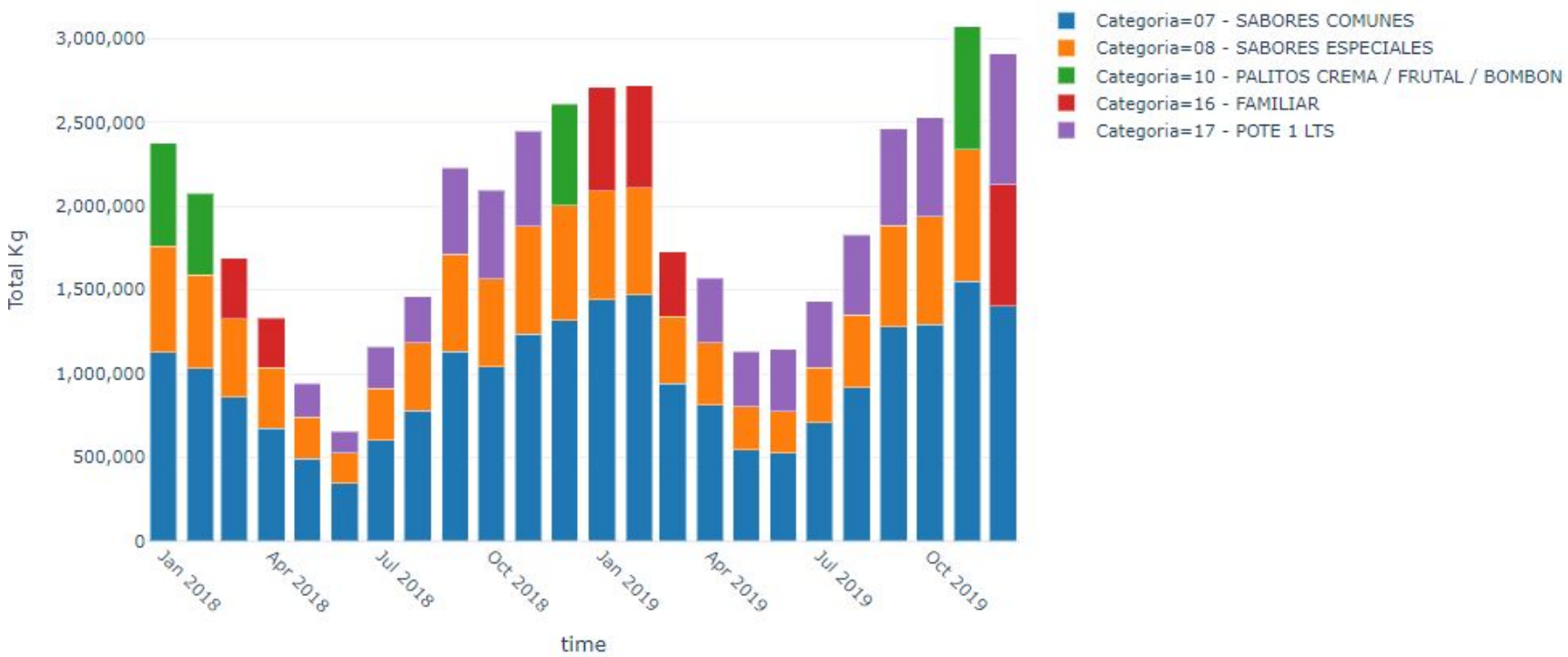


Variables de interés: Localidad



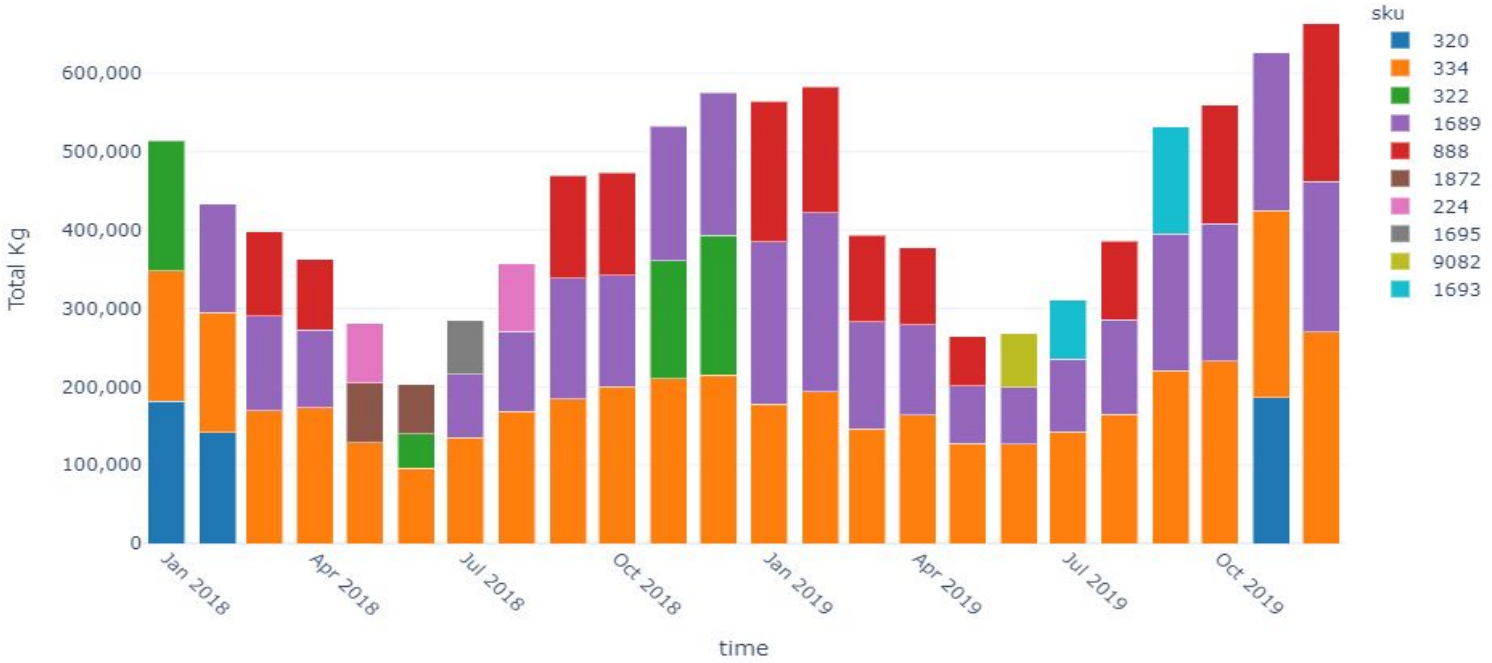
Mensualmente el top 3 de pedidos por localidad tiene siempre a Córdoba como mayor demandante pero las demás posiciones varían (total de 600 localidades)

Variables de interés: Categorías



Variables de interés: Producto

- ✔ El sku 334 (13 - IMPULSIVOS) en general domina a lo largo de todo el año, incluso en invierno. Otros productos tienen mayor estacionalidad como el 1689(07 - SABORES COMUNES)
- ✔ Hay productos que logran entrar entre los tres más vendidos solo en verano, como el 320 (10-PALITOS HELADOS) y otros solo en invierno, como el 1872 (01-FRIZZIO).



Resultados del Análisis Exploratorio

- ✔ **Asimetría**: la distribución de los pedidos es en general asimétrica teniendo una mayor densidad de pedidos pequeños

- ✔ **Estacionalidad**: la demanda en los meses de verano aumenta mientras que para los meses de invierno la demanda cae. También se percibe que en promedio la demanda para el año 2019 es mayor a la del 2018.

- ✔ **Territorialidad**: pedidos bajos con mayor frecuencia para Córdoba mientras que existen pedidos de gran magnitud con menor frecuencia para Tierra del Fuego.

Machine Learning

1. Introducción y Objetivos
2. Descripción Dataset
3. Análisis Exploratorio y Curación de Datos
4. Machine Learning

Machine Learning

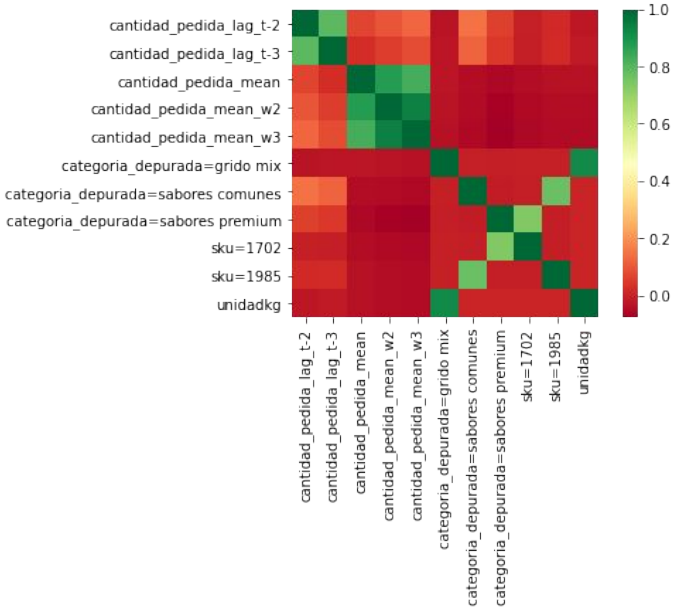
Preprocesamiento:

- Se redujo la cardinalidad de las variables ‘sku’ y ‘ubicación’ agrupando las categorías más comunes en ‘otros’
- Se filtraron outliers de la variable ‘cantidad_pedida’ y el resto de las variables numéricas se mantuvieron igual
- Las variables categóricas se codificaron con one-hot encoding
- Se enriquecio el dataset con nuevas variables para transformar en un problema de regresión:
  - Variables lag t-n que representan la sumatoria de los valores de ‘cantidad_pedida’, para la variable ‘Ubicacion’ y ‘sku’ agrupados por mes para “n” periodos anteriores. Se seleccionó n = 3 para capturar la información del cambio de estación
  - Medias móviles simples para la columna cantidad_pedida. Se utilizó una ventana de 2 y 3 periodos anteriores.
  - Diferencia entre la variable dependiente y su valor (shift) anterior, para eliminar la tendencia.

Machine Learning

Preprocesamiento:

- Se eliminaron variables no relevantes. Para esto se hizo un análisis de multicolinealidad y correlación



- Lags y medias móviles muy relacionados entre sí. Es de esperarse por construcción, por lo que no se eliminaron.
- Categorias corelacionadas con 'unidadkg'y skus se eliminan.
- Del análisis de multicolinealidad se obtuvo que varias categorías, marcas, presentaciones y id_proveedor aportaban información redundante con la columna sku así que también se eliminaron.

Machine Learning

Preprocesamiento:

- Se filtró por la provincia y localidad de Córdoba.
- División test/train: se utilizó un 20% de los datos disponibles como conjunto de test y el restante 80% para entrenamiento. La división de train y test debe respetar la línea de tiempo.
- Dentro del 80% de train, no es posible utilizar K-Fold Cross Validation dado que esta forma de validación cruzada no respeta el orden de las observaciones. Por eso se utilizó la función TimeSeriesSplit() con n_splits=10.
- Se utilizaron algoritmos de regresión basados en árboles (Random Forest, XGBoost, LGBM), así que no fue necesario normalizar o estandarizar la data.

Machine Learning

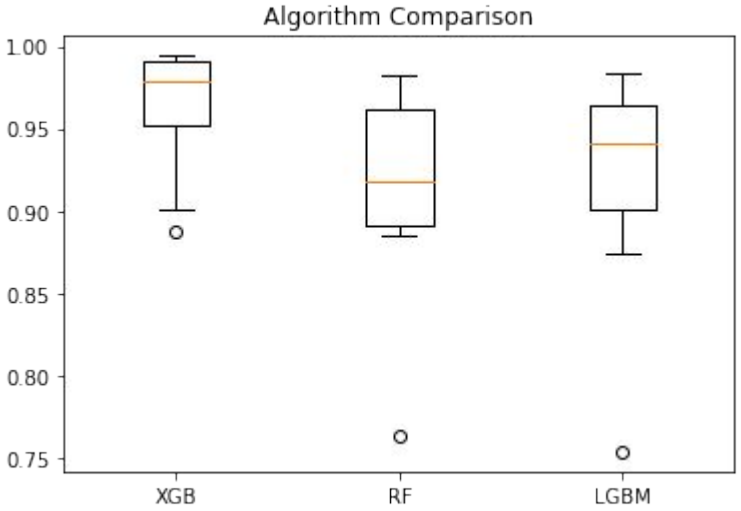
Entrenamiento:

- Se investigó, para cada modelo, los hiperparámetros más importantes y su posible rango de valor y con esta información se procedió a armar grillas de parámetros.
- Se llamó a la función GridSearch(), pasándole como argumentos el modelo, la grilla correspondiente a ese modelo, la métrica respecto a la cual optimizar los hiperparámetros, en este caso el r2-score, y los folds obtenidos de TimeSeriesSplit().
- Finalmente se evaluaron los resultados y se guardó el modelo con los hipérparametros que maximizan la métrica.

Machine Learning

Resultados:

La mayoría de los modelos tuvieron una muy buena performance, con un r2-score de más de 0.9. Sin embargo, el XGBRegressor fue el que tuvo los mejores resultados.

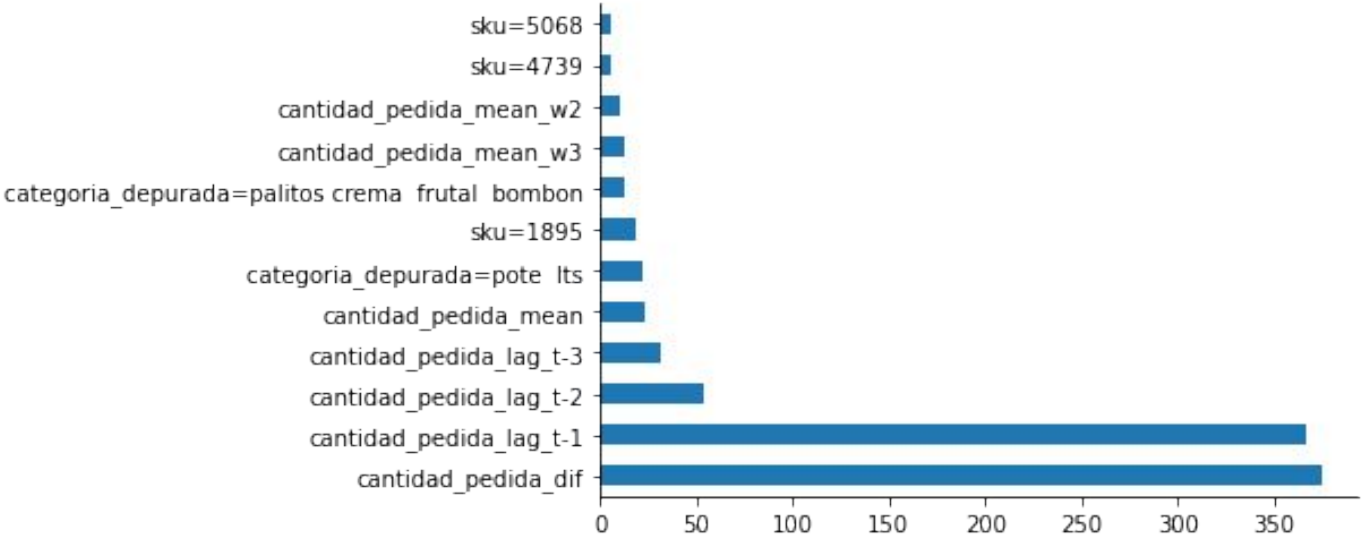


modelo	rmse	mape	r2
XGB	11.547643	0.201504	0.987865
RF	22.110115	0.266866	0.955514
LGBM	24.496418	0.258762	0.945393

Machine Learning

Resultados:

Feature Importance



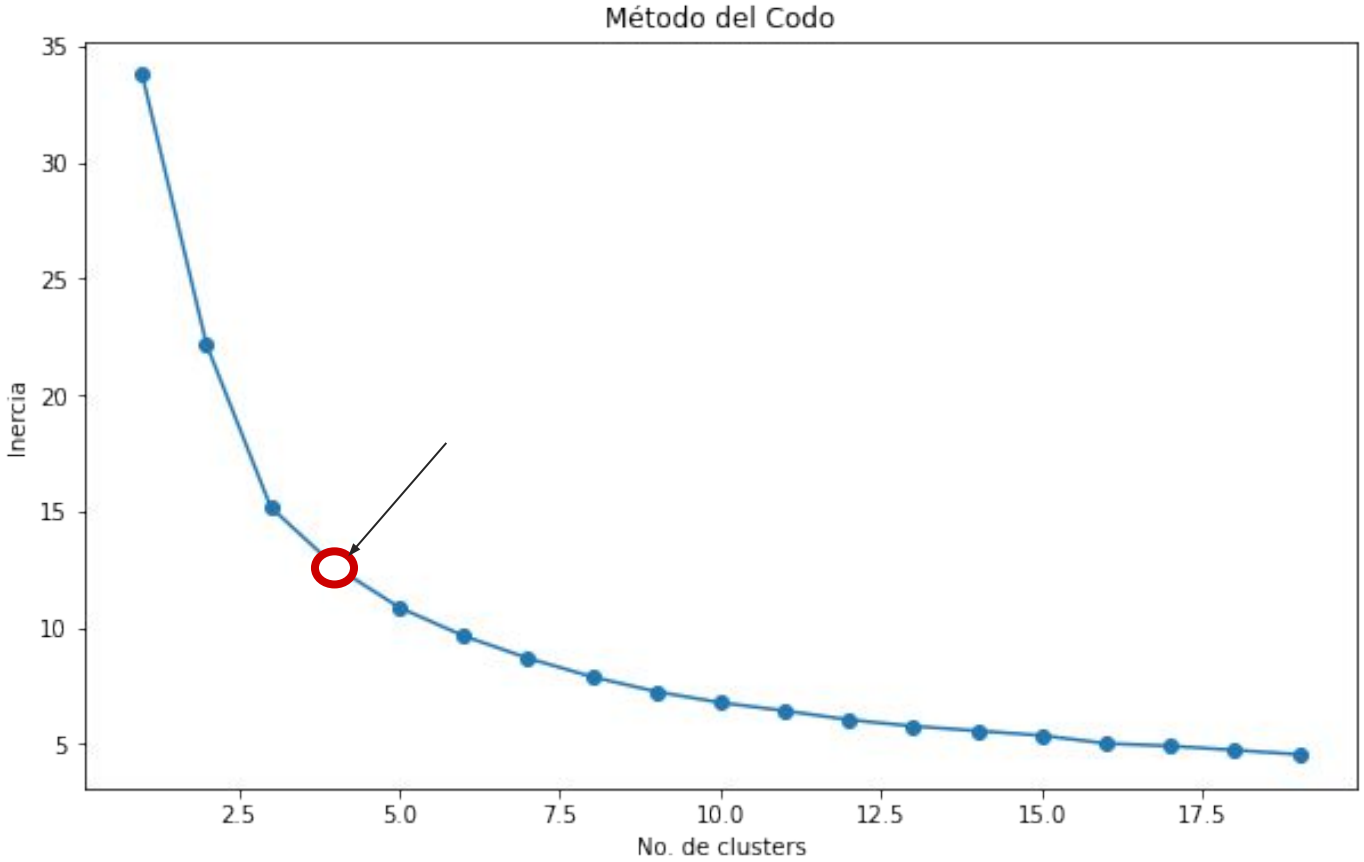
Clusterización

1. Introducción y Objetivos
2. Descripción Dataset
3. Análisis Exploratorio y Curación de Datos
4. Machine Learning
5. Clusterización

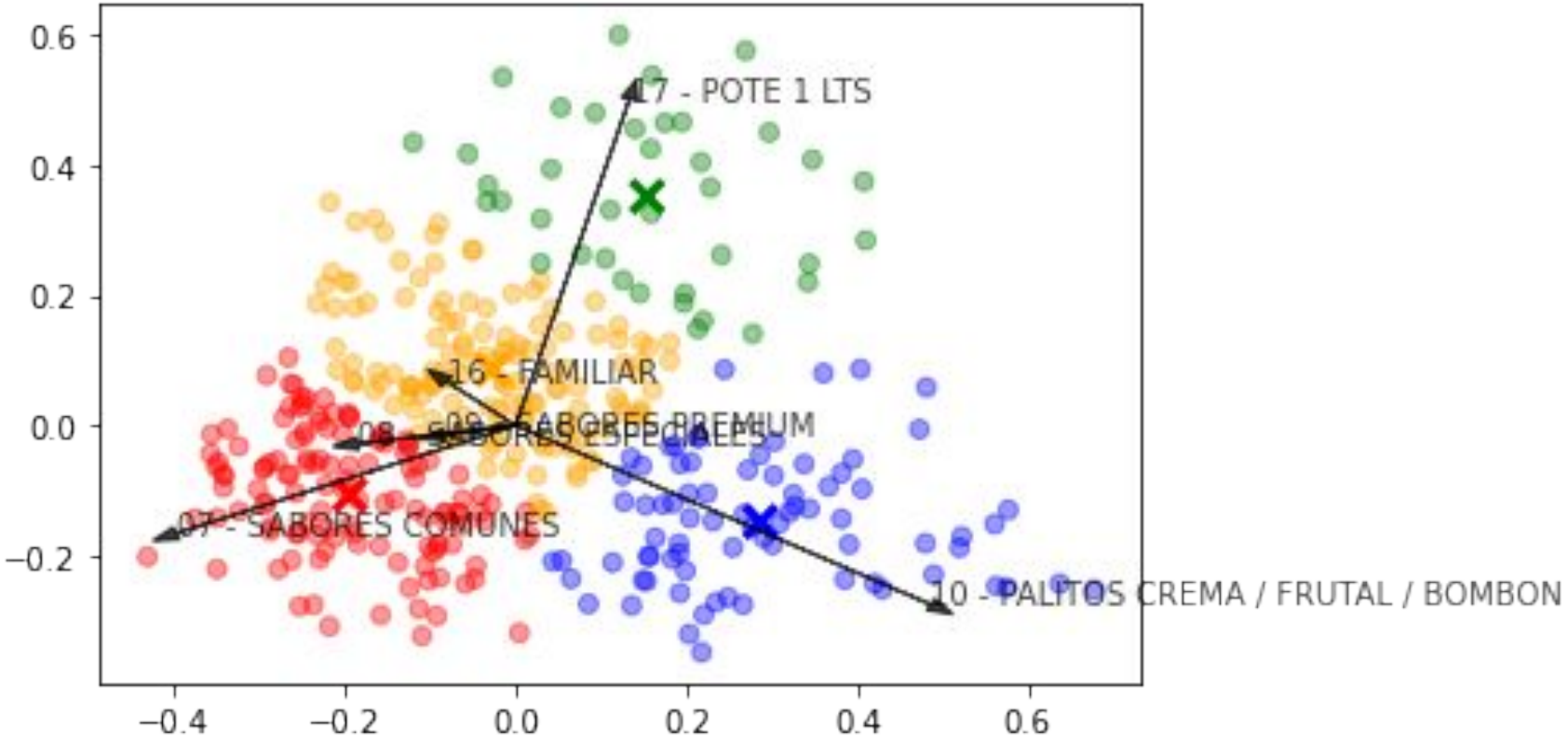
Clusterización

- Objetivo: identificar grupos de productos que se suelen vender en forma conjunta
- Se trabajó sobre una versión reducida del dataset original que contenía:
  - 6 categorias de productos mas vendidos (07 - SABORES COMUNES; 08 - SABORES ESPECIALES; 17 - POTE 1 LTS; 16 - FAMILIAR; 10 - PALITOS CREMA / FRUTAL / BOMBON; 09 - SABORES PREMIUM)
  - 3 provincias que más venden (Buenos Aires, Córdoba, Santa Fe)
- Pretratamiento: curado y normalizado
- Algoritmo: K-means
- Ajuste de hiperparámetros: método del codo

Clusterización



Clusterización



Desafíos y Conclusiones

1. Introducción y Objetivos
2. Descripción Dataset
3. Análisis Exploratorio y Curación de Datos
4. Machine Learning
5. Clusterización
6. Desafíos y Conclusiones

Desafíos y Conclusiones

Algunas conclusiones generales…

- Cualitativamente la demanda que analizamos de alimentos congelados tiene como características la asimetría, territorialidad y estacionalidad.
- En general, la mayoría de los modelos tuvieron una muy buena performance, con un r2-score de más de 0.9
- Los modelos que se seleccionaron finalmente incluyen un Regressión Tree y un Extreme Gradient Boosting Regressor.
- Los patrones de clustering que obtuvimos tuvieron un cierto paralelismo con las categorías del los productos ya existentes.
- Sería interesante explorar las posibilidades de relacionar con contenido de las materias optativas y como traducir un proyecto de ML a un producto comercialmente interesante.

Desafíos y Conclusiones

Dentro de lo que consideramos fue desafiante del proceso de mentorías podemos mencionar...

- Dificultades:
  - Tamaño del data set
  - Manejo del tiempo

- Cuestiones valiosas:
  - Series de tiempo
  - Trabajo con grupos en paralelo
  - Enriquecimiento del dataset

- Cuestiones a explorar:
  - El dataset del año 2020, datos de pandemia
  - implementación de un producto a campo
  - redundancia de algunas tareas

¡Muchas gracias por su atención!

Se pueden encontrar los notebooks e informes más detallados en github
https://github.com/sofianieva/prediccion_demanda_grupo_2