Faltante de Mercadería en Góndola (FMG): Análisis Mercado General y Panificados en la Argentina

Abstract: La necesidad de obtener modelos predictores que sirvan para la toma de decisiones, se expande hacia distintos horizontes y posibilidades como lo son los mercados de bienes y servicios. El objetivo de este paper, es determinar un modelo predictor para quiebres de productos panificados en góndolas y obtener información del mercado que sirva para establecer planes comerciales. Para esto, se utiliza modelos clasificadores de aprendizaje supervisado. Los resultados arrojan una relación entre las variables que debe ser considerada.

Supervisión:
Mg. Ing. Palazzo, Martín
Mg. Ing. Nicolas Aguirre
Mg. Ing. Agustín Velázquez
Maestría en Sistemas Complejos – UTN.BA-UTT
Mauro Lucini
Ingeniería Industrial – UTN.BA

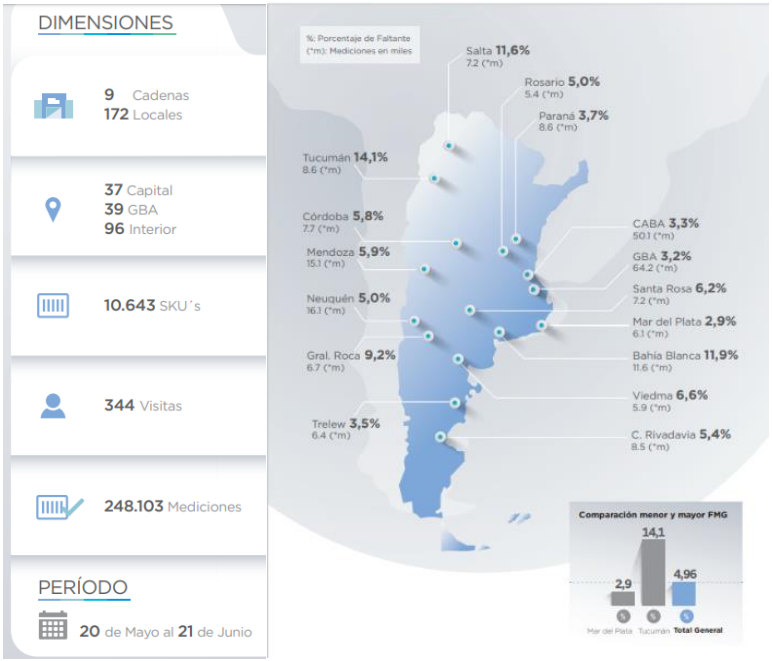
Realización:
Strauss, David
Bachur Solari, Alejandro
Castaño, Gabriel
Ingeniería Industrial –UTN.BA

1. Introducción

El siguiente trabajo se realizó con la finalidad de predecir los quiebres de góndolas de productos panificados en el ámbito de la República Argentina. Este estudio está destinado a la Panificadora de Córdoba PANNI la cual está dimensionando su estructura comercial y le será muy útil conocer la situación actual y futura del mercado.

El dataset utilizado en el presente trabajo fue brindado por GS1 Argentina. Anualmente, la asociación mencionada realiza el estudio de Faltante de Mercadería en Góndola (FMG) en las principales ciudades del país. El Estudio FMG permite conocer la disponibilidad de los productos en las góndolas y conocer las causas que originaron los quiebres de stock.

A partir de la información obtenida por dicho estudio en la República Argentina en el año 2019 en 14 cadenas, se espera encontrar información de los 13.000 productos a nivel país, y compararlo con los productos panificados en la zona central para brindar información de la plaza al proyecto de inversión PANNI. Además, pretende modelizarse para productos panificados el faltante en base al stock, la rotación, localidad, la cadena y al tipo de distribución: directa (N), Cross-Docking (X) o centro de distribución (S).



2. Descripción del dataset

Los datos iniciales del 2019 componen de 326.928 mediciones (samples) y 30 variables (features). Debido a que los datos brindados pertenecen a entidades privadas, se debe mantener confidencialidad de los datos. Para ello se procedió a realizar un enmascarado de la información de las cadenas de supermercados, proveedores, marcas y productos.

```
marcas=np.unique(gondolas.MARCA)
Buscarvmarca = pd.DataFrame(marcas, columns = ['Marcas'])
auxiliar2 = int(np.shape(Buscarvmarca)[0])
mascaramarca= [0]*auxiliar2
for i in range(0,auxiliar2):
    mascaramarca[i]= ('Marca'+' '+str(i))
mascaramarca
Buscarvmarca['mascara']= mascaramarca
Buscarvmarca.head()
```

Las features utilizadas son:

- **IDRENGLON:** número identificatorio e irrepetible de cada medición (campo key).
- **IDMUESTRA:** número identificatorio de la visita al local; agrupa un conjunto de mediciones.
- **CODIGO_EAN:** número identificatorio del producto, está relacionado con el código de barras de 13 dígitos. Fue enmascarado.
- **MARCA**: del producto medido. Fue enmascarado.
- **ID_VALIDO:** es el código de motivo asociado al faltante.
- **STOCK_CD:** es la cantidad de stock del producto en el Centro de Distribución de la cadena de supermercado.
- **PROVEEDOR:** es el nombre del proveedor del producto. Fue enmascarado.
- **CADENA:** es el nombre de la cadena de supermercado. Fue enmascarado.
- **ZONA:** es el nombre de la ciudad donde se hizo la medición.
- **C_TAMANIO:** es el formato del local: Hiper, Super, Mini, Proxi o Farma.
- **IDROTACION:** hace referencia al tipo de rotación del producto: Alta (A), Media (B) o Baja (C).
- **ENTREGA:** hace referencia al tipo de entrega: Centralizada (por CD de la cadena) o Descentralizada (directa del proveedor).
- **IDALMACEN:** es el tipo de almacén de la cadena: entrega directa proveedor (N), Cross-Docking (X) o Centro de Distribución (S).
- **FECHA_RELGS1:** fecha de la medición.
- **ANIO:** año del estudio.
- **DIASEMANAL:** corresponde al día de la semana de la medición: lunes, martes, etc.
- **QUIEBRE:** de góndola: Sí (S) o NO (N).
- **CAT2:** categoriza a los productos en familias.
- **CAT3:** categoriza los productos en: alimentos, bebidas, higiene personal, cuidado del hogar o salud.
- **SUB_CAT3:** es una subcategoría de la CAT3: alimentos secos o perecederos, bebidas con o sin alcohol, etc.

3. Análisis exploratorio de Datos

Se trabajó con el lenguaje de programación Python 3 desde las plataformas de Jupyter para llevar a cabo el análisis de los datos.

A partir de estas variables se realizó un pre-processing:

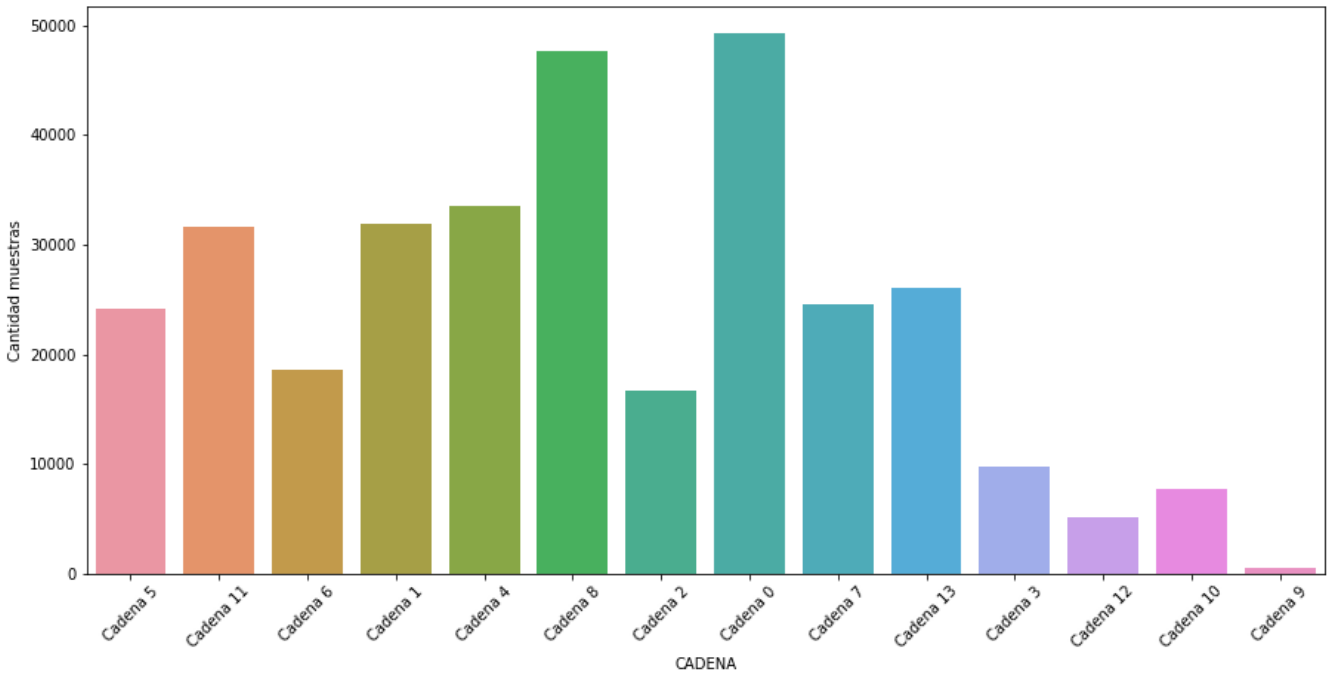
- se eliminaron features que no hacían al análisis; como ID_MOT_CADENA, ID_MOT_PROV, ID_SUBMOTIVO, TAMANIO, IDGRUPO, CATEGORIA, ORIGEN.
- se transformaron tipos de valores de features; como los valores de FECHA_RELGS1.
- se transformaron valores de features; como en la feature ANIO o QUIEBRE.
- Se procesaron los NULLS tanto en filas como en columnas.

Se hizo un análisis exploratorio de datos. La información más relevante se detalla a continuación:

Cantidad de Mediciones por Proveedor:

PROVEEDOR	IDRENGLON	Porcentaje
Proveedor 46	30411	9.302048
Proveedor 626	24270	7.423653
Proveedor 443	22940	7.016836
Proveedor 442	15160	4.637107
Proveedor 121	10157	3.106800

Cantidad de Mediciones por Cadena:



Cantidad de Mediciones por categoría:

CAT2	IDRENGLON	Porcentaje
GALLETITAS DULCES	13382	4.093256
CERVEZA	12373	3.784625
JUGOS EN POLVO	12133	3.711215
FIDEOS SECOS	9371	2.866380
GASEOSAS	7674	2.347306

Top 5 Quiebres por Zona:

ZONA	QUIEBRE
CABA	5422
GBA	3994
CORDOBA	3733
BAHIA BLANCA	2094
NEUQUEN	1697

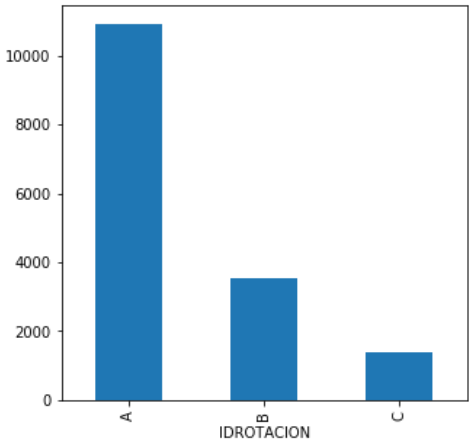
Cantidad de quiebres por Tamaño del mercado:

C_TAMANIO	QUIEBRE
HIPER A	7572
SUPER	6683
MINI	4479
FARMA	3990
HIPER B	3331

Cantidad de Quiebres por rotación del producto:

IDROTACION	IDRENGLON	Rot_quiebre	Quiebre/Total
A	210915	10906	5.170803
B	51838	3528	6.805818
C	15029	1405	9.348593

Siendo IDRENGLON la cantidad de mediciones.



Top 5 Quiebres por categoría

CAT2	QUIEBRE
CERVEZA	2667
FIDEOS SECOS	915
CREMA CORPORAL Y MANOS	912
GALLETITAS DULCES	833
JUGOS EN POLVO	750

Finalizado el análisis genérico, se procedió a filtrar el dataset acotando el EDA a aquellos productos de la categoría panificados (feature CAT2) y se obtuvo como conclusión más relevante:

Sobre el total de muestras el 8.76% son muestras con quiebre. En cambio, sobre la categoría panificados el 19.13% de las muestras están quebradas.

- Respecto a las zonas del país con mayores porcentajes de quiebres obtuvimos:

ZONA	IDRENGLON	Zonas_quiebre	Quiebre/Total
SANTA ROSA	28	23.0	82.142857
SALTA	46	22.0	47.826087
VIEDMA	28	12.0	42.857143
TUCUMAN	70	29.0	41.428571
CORDOBA	184	62.0	33.695652
NEUQUEN	60	19.0	31.666667
GRAL. ROCA	52	14.0	26.923077
C RIVADAVIA	34	7.0	20.588235
GBA	426	82.0	19.248826
RAFAELA	132	24.0	18.181818
PARANA	75	7.0	9.333333
BAHIA BLANCA	132	9.0	6.818182
CABA	362	24.0	6.629834
MENDOZA	90	2.0	2.222222
ROSARIO	27	NaN	NaN
MAR DEL PLATA	10	NaN	NaN

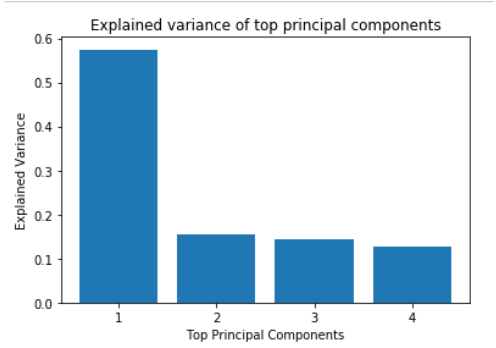
Toda esta información resulta clave para la definición del plan comercial de PANNI. A través de un Merge con ID_VALIDO y una Tabla de motivos tabulados, se agruparon y mensuraron las causas de quiebre en panificados:

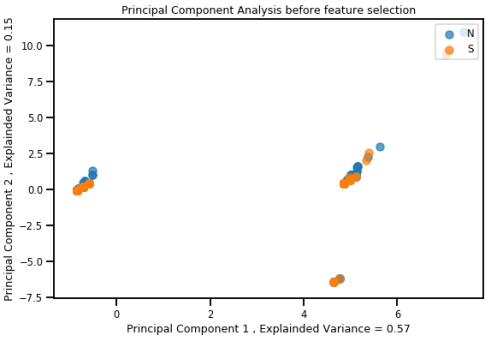
	IDRENGLON	Porcentaje_total	Detalle	fechaAlta
0	78	23.214286	101	Falso stock positivo en el local
1	47	13.988095	103	Mercadería rota/vencida
2	44	13.095238	407	Pedido pendiente de entrega del proveedor al l...
3	32	9.523810	102	El local no realizó el pedido
4	30	8.928571	400	El proveedor no entregó (general)
5	25	7.440476	108	Faltante por reposición INTERNA
6	22	6.547619	450	Otros problemas con el proveedor
7	17	5.059524	250	Otros problemas del CD
8	13	3.869048	111	Falso stock negativo en el local
9	9	2.678571	107	Faltante por reposición EXTERNA
10	9	2.678571	207	Orden de compra pendiente de entrega a la tienda

4. Machine Learning Metogología

Determinación de los quiebres a través de modelos de aprendizaje supervisado. Se utiliza modelos de clasificación para la predicción de quiebres de stocks en el segmento de panificados. Las features de entrenamiento seleccionadas para el estudio del comportamiento de quiebre (Y: “quiebra” o “no quiebra”) son: el tipo de entrega, la rotación, stock en el centro de distribución y el tipo de almacén de

la cadena. Se generan las muestras de entrenamiento y testeo y una primera visualización con un PCA, para tener una idea rápida de cómo vienen los datos, que a simple vista son difícilmente separables.





Procedemos entonces a ejecutar los modelos de aprendizaje. Se generó un modelo SVM que asignó todas las muestras de testeo a "no quiebre". Además se estudió utilizando KNN, y se observó qué también tiene dificultades para predecir las muestras con quiebre.

La consigna es entonces, mejorar los modelos obtenidos. Lo primero que se hizo fue incorporar más datos (se generaron más features con la Zona y la Cadena). Se volvió a probar un KNN, que mejoró su accuracy respecto al predictor anterior. Sin convencer el resultado del modelo, se pasó a utilizar Logistic Regression, que arrojó un resultado similar (leve menor accuracy).

Entonces se pensó en balancear las muestras, y se probó primero un Logistic Regression balanceado, pero sin buenos resultados.

Finalmente, se intentó como última técnica *rellenar* el dataset con muestras ficticias de la etiqueta de menor presencia y eliminar las que "sobran" del tipo de etiqueta con quiebre. Se volvió a probar utilizando KNN.

5. Resultados

Cabe mencionar la importancia que ostentaba no solamente tener una elevada precisión o exactitud en las predicciones generales, sino también encontrar un modelo que predijera con una confiabilidad aceptable los quiebres de stock. En la búsqueda de este resultado, se obtuvo en primer lugar un Support Vector Machine y a pesar de que el accuracy arrojado es de 0,795, el modelo no clasificaba productos con quiebre (Recall "quiebre" = 0%).

En su lugar, el primer modelo KNN si bien posee un accuracy similar (80%), comienza a predecir la presencia de productos "con quiebre", con un 27% de Recall de quiebre. Su respectiva Confusion Matrix (CM):

```
     0   1
0[ 394  25]
1[ 79   29]
```

Siendo el KNN, más preciso que su antecesor, se procedió a probarlo nuevamente, ahora incorporando mayor información al dataset. Lo cual arrojó un accuracy del 84% y una detección del 35% de los productos quebrados.

Por su parte, el modelo de Logistic Regression obtiene resultados similares (83% y 32% Recall clase 1). No siendo satisfactorio estos resultados se los compara con modelos balanceados.

De esta forma, se balancea un Logistic Regresion, modelo en el cual se desploma considerablemente el accuracy (65% , pero se obtiene una precisión de detección de las muestras quebradas del 91%). Teniendo este último una elevada detección de la clase, pero incluyendo demasiadas muestras de la otra, lo cual implica un costo altísimo. CM:

```
    0    1
0[254  172]
1[  9   92]
```

Por último, los modelos balanceados con muestras adicionales, KNN por un lado y LR por el otro, nos dan resultados similares:

KNN ➔ Accuracy 81,4% y CM (50% de Recall clase con quiebre):

```
    0    1
0[379   47]
1[ 51   50]
```

LR➔ Accuracy 81,4% y CM (60% de Recall clase con quiebre):

```
   0    1
0[368   58]
1[ 40   61]
```

6. Conclusiones

Se observa una relación interesante en la rotación del producto y la cantidad de quiebres según el análisis exploratorio de datos.

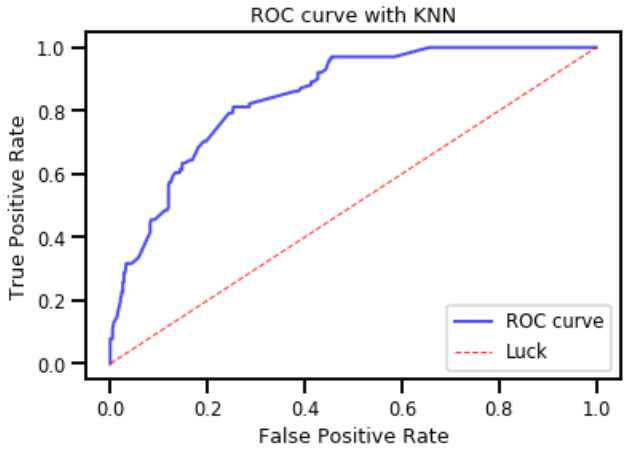
Como los datos son difícilmente separables y el 80% de las muestras son de una etiqueta, los modelos aprendidos tienen dificultades para poder diferenciar las samples. Esta situación se refleja notablemente en el primer SVM y KNN (sin generar datos).

Como segunda conclusión, al otorgarle a las muestras de entrenamiento mayor información en las features adicionadas, el algoritmo incrementa su confiabilidad. El segundo KNN (con generación de datos) detecta el 35% de los quiebres y el accuracy es de 84%. Es así como entra en juego, el balanceo de muestras.

El accuracy Logistic Regresión balanceado cae a un 66%, porque si bien detecta prácticamente todas las muestras que estaban quebradas, lo hace con un costo de bajar la precisión, ya que empieza a decir que la mitad de las muestras tienen una etiqueta y la otra mitad otra (como si las posibilidades fueran 50 y 50).

Para el tercer KNN, última técnica (con generación y balanceo de muestras) encuentra la mitad de las muestras quebradas. Si lo comparamos con un Logistic Regression que tiene similar accuracy (0,814), pero encuentra el 61% de las muestras quebradas. Lo cual mejora sustancialmente la predictibilidad de las muestras con quiebre, sin renunciar a la exactitud general del modelo.

Este último es el que mejor se adapta al comportamiento de los datos dado el objetivo trazado al comenzar el estudio, y se obtuvo que el ROC bajo la curva es 84. Se considera entonces que un modelo aceptable.



7. Referencias

1.https://www.gs1.org.ar/documentos/FMG/FMG_2019.pdf

2.https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/

3.Patter Recognition and Machine Learning, 2006, Christopher Bishop.

4.https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/