Aplicación de Machine Learning a Predicción de Demanda

Segundo Informe de Análisis y Curación

4 DE JULIO DE 2021

Integrantes:
Gutiérrez Montecino, Denise
Nieva, Sofía
Rodríguez, Alfredo Manuel

Contenido

1 INTRODUCCIÓN

En el siguiente informe se detallan los resultados obtenidos del análisis y curación de la base de datos suministrada como parte de la Mentoría ***Predicciones de Demanda de Producto*** de la Diplomatura de Ciencia de Datos de la Facultad de Matemáticas, Astronomía y Física (FaMAF) de la Universidad Nacional de Córdoba, Argentina.

El conjunto de datos de partida contiene productos vendidos de los últimos 5 años de una compañía de venta de alimentos congelados en distintos países de la región. El objetivo final de la mentoría es poder predecir la demanda de los productos elaborados en los centros de elaboración mes a mes en los diferentes países y zonas en donde opera esta compañía.

El presente entregable contiene documentación del proceso de curación de la base de datos que va de la mano con la verificación de consistencia de la información. También se realiza nuevamente un análisis exploratorio de la base de datos, teniendo en cuenta lo analizado en el informe anterior. El objetivo de este informe es obtener una base de datos que sirva de entrada para los modelos de aprendizaje automático con posibles iteraciones futuras.

2 RESPUESTAS DEL PRÁCTICO

En esta sección se desarrollarán las consignas dadas para el informe con sus respectivas observaciones y comentarios de los resultados obtenidos así como también de las decisiones que se fueron realizando, si fuese el caso. A continuación se describe cada una de ellas.

1. Importación Datos

a. Verificación de Inexistencia de Problemas en la Importación

En esta sección se importan las bases de datos y se analizan cada una de ellas. Se inspeccionaron los tipos de datos que tienen cada una de las tablas, el tamaño y las columnas que posee.

En caso de que exista algún tipo de corrección informado por la contraparte que proporciona las bases de datos se debe realizar en este momento. Se corrigieron los precios de los siguientes productos: 3170, 3171, 12150, 3095 y 22660.

b. Asegurar la existencia de ID’s o claves únicas

En esta sección se chequea que no existan datos duplicados y que las claves sean únicas. No se encontraron datos duplicados para los productos ni para los puntos de venta. En otras palabras, existe un único sku para cada producto y un único id para cada punto de venta.

2. Pasos Importantes para Limpieza de Datos

a. Etiquetas de variables/columnas y problemas de codificación/encoding

Se realizan dos análisis, uno a nivel registros de la base de datos y otro enfocado a los nombres de las variables de la misma.

Para el primer caso se observa que los registros no tienen la misma codificación. Sin embargo, en los puntos que se desarrollarán luego se “limpian” aquellos registros de las columnas que serán utilizadas para el One Hot Encoding (ver punto 5.1).

En el segundo caso, se observa que todas las variables poseen caracteres entre A-Z, 0-9 y “-”.

b. Tratamiento de valores faltantes

Se analizan cada una de las bases de datos en busca de posibles valores faltantes.

En primer lugar se analiza el dataset de productos. No se identifican valores nulos (o NaN) aunque se observa que existen 2804 productos con marca “Sin definir” y 15 productos que poseen un string vacío en presentación. Estas últimas representan un 0.0034% de un total de 4532 productos distintos, en el cual 10 de estas poseen marca “Sin definir” mientras que los otros 5 sí poseen una marca determinada. Considerando que los datos faltantes son muy pocos se podrían descartar o agrupar en una categoría “Sin presentación”.

En segundo lugar se procede a analizar el dataset de categorías. No se identifican valores nulos (o NaN) aunque se observa que existen 6 registros que poseen nombre de categoría “VACIO” o “vacia”. Estas representan un 0.0682% de un total de 88 categorías distintas. En este caso no será necesario imputar los valores faltantes dado que no existe ningún producto asignado a esa categoría.

Siguiendo con el análisis se estudia el dataset de provincias, se concluye que no posee datos faltantes pero si se hallan valores duplicados. No hay datos faltantes pero hay nombres de provincias repetidas. Una suposición es que pertenecen a países distintos. De todas maneras, no será necesaria la imputación dado que en el dataset de pdv aparecen con un sólo id_ Provincia, probablemente porque hay un sólo país en el dataset.

Por último, se procede a analizar los datasets de puntos de ventas, ventas, localidades y países en el que no se encuentran datos faltantes de ninguna clase (ni valores nulos, vacíos, etc).

c. Codificación de variables categóricas. ¿Aplica codificar variables categóricas?

Dado que el objetivo de este informe es preparar un dataset lo suficientemente óptimo para poder realizar a futuro un modelo de Aprendizaje Automáticos es sumamente importante y necesario que las variables categóricas se codifiquen. Para poder realizar el mismo existen distintos métodos, el más utilizado corresponde al One Hot Encoding que se desarrollará más adelante.

d. Verificación de consistencia de datos

Este punto se desarrollará en el punto 4 inciso a del informe.

e. Identificar y documentar valores atípicos/outliers.

Calcular estadísticas y eliminar los outliers utilizando el método que considere pertinentes (z-score, z-score modificado o IRQ) justificando y documentando la elección realizada. Realizar el análisis sobre la columna cantidad_pedida

Dado que la distribución de la variable ***‘cantidad_pedida’*** es asimétrica, aún considerándolo en escala logarítmica, se deben utilizar métodos no paramétricos para la detección de outliers o valores atípicos. En la figura a continuación se muestra la distribución de la variable en cuestión.

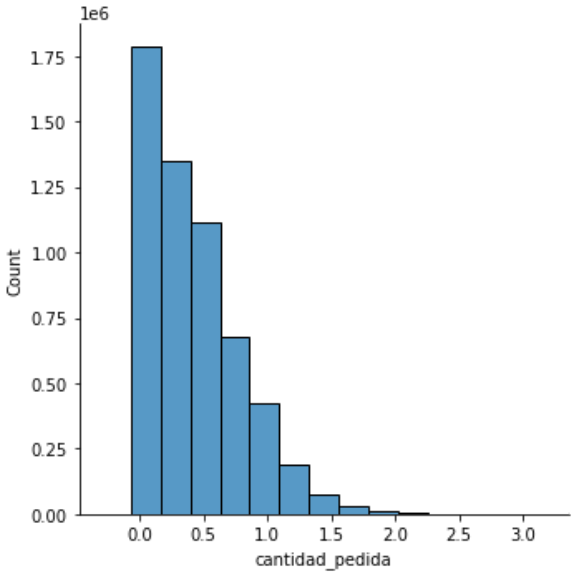

Figura 1 - Distribución logarítmica de la variable cantidad_pedida

Algunos de los métodos no paramétricos son el IQR, el uso de los percentiles, entre otros. Tras una serie de iteraciones y pruebas de cuál es el método más adecuado para extraer se seleccionó el percentil 0.995 y 0.005 dado que las demás opciones eliminaban demasiados registros. A continuación se observa en la figura los boxplot para la variable cantidad_pedida sin eliminar los outliers y cómo queda luego de la eliminación de los mismos.

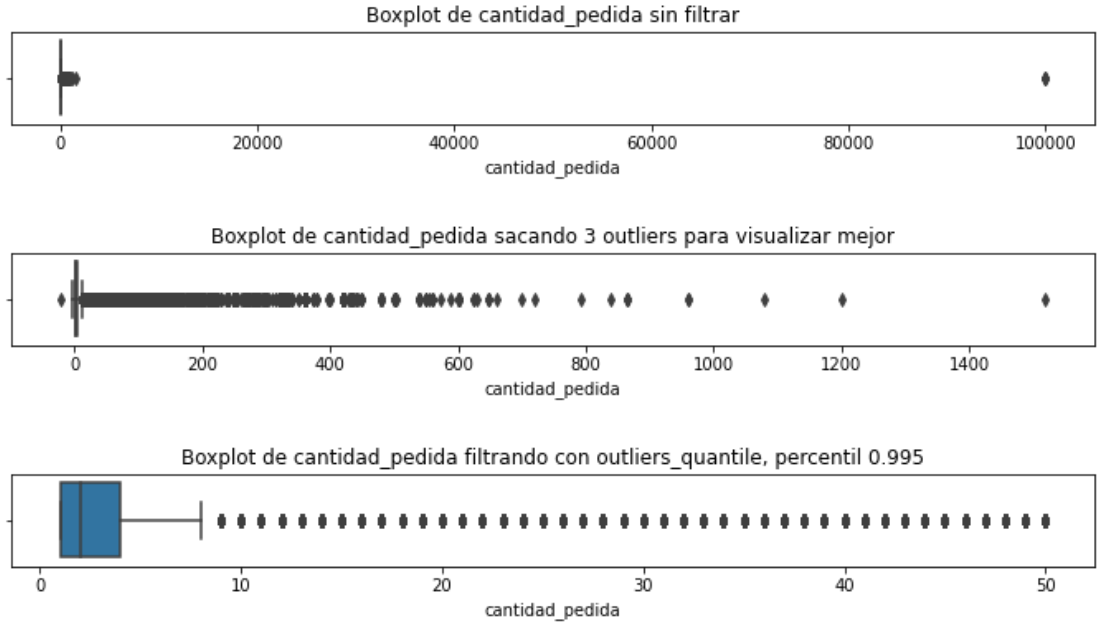

Figura 2 - Boxplot de la variable ***'cantidad_pedida'*** con outliers y sin outliers

3. Pasos Opcionales para Limpieza de Datos - Deseables

Se procedió a eliminar algunas de las columnas que se consideran irrelevantes para el análisis. A cotninuación se explica brevemente estas decisiones:

- ***'dia'***: Se eliminó ***'día'*** dado que el objetivo es predecir la demanda de los productos a nivel mensual.

- ***'hora'***: Al igual que la variable anterior, se eliminó dado que el objetivo es predecir la demanda de los productos a nivel mensual.
- ***'descripcion':*** podría considerarse una variable redundante dado que se tiene la variable ***'sku'*** del producto, siendo esta última una mejor alternativa por ser más sencilla de procesar al ser un número.
- ***'id_categoria':*** podría considerarse una variable redundante dado que se tiene la variable '***categoria'*** del producto. En este caso se mantiene esta última dado que es necesaria para el desarrollo de los ítems posteriores.
- ***'unidadcm3':*** esta variable no es utilizada para el análisis dado que se opta por las variables de ***'unidadkg'*** y ***'cantidad_pedida'***, o como se verá más adelante, la variable ***'totalkg'*** resultante del producto de ambas variables mencionadas.
- ***'id_punto_venta'*** y ***'Punto_Venta'***: se eliminan ambas variables dado que el objetivo del estudio es predecir los productos por localidad, por lo que no será necesario considerar los puntos de ventas.
- ***'id_Localidad'***, ***'id_Provincia'***, ***'id_Pais'***, ***'Localidad'*** y ***'Pais':*** se generó una nueva variable denominada ***'ubicacion'*** resultante de la combinación de ***'Localidad'*** y ***'Provincia'*** por lo que se pueden eliminar las otras variables. Por el momento se conserva la variable ***'Provincia'*** dado que se requiere para el desarrollo de incisos posteriores pero después se podría eliminar.

4. Análisis Complementario de algunas Features

a. Al generarse los joins entre todas las fuentes de datos, ¿qué información quedó excluida?

De la primera unión entre productos y categorías se obtuvo un nuevo dataset denominado producto_categoria. Para el caso del dataset de productos se conservó la totalidad de los registros. En el caso del dataset de categorías, existen 13 que ya no se encuentran en el nuevo dataset de producto_categoria. La razón es que no existen productos con estas categorías, algunas de las categorías que ya no se observan son las categorías de "VACIO" y "vacia". La exclusión de esta información no genera ningún tipo de inconveniente dado que el objetivo del estudio es el análisis a nivel producto y dado que estas categorías no estaban asignadas a ningún producto no serán necesarias en este estudio.

De la segunda unión entre producto_categoria y ventas_filtrado se obtuvo un nuevo dataset denominado ventas_producto. Para el caso del dataset de ventas_filtrado se conservó la totalidad de los registros, aunque se observa que de un total de 4352 de sku sólo quedaron 588 sku. Esto no generaría inconvenientes dado que la ausencia de estos productos seguramente se deba a que no se registraron ventas en la base de ventas_filtrado para los mismos. Otra observación importante para destacar es el caso del sku 9282 que figura en el nuevo dataset que posee información de ventas pero no de las columnas provenientes de producto_categoria (dado que no existe en este último).

En la tercera unión entre pdv y localidad se obtuvo un nuevo dataset denominado pdv_ubicacion. Se observa que existen 164 localidades que ya no se encuentran en el nuevo dataset, esto puede ser debido a que las mismas no poseen puntos de ventas.

En la cuarta unión entre pdv_ubicacion y provincia se obtuvo un nuevo dataset denominado pdv_ubicacion. Se observa que existen 53 provincias que ya no se encuentran en el nuevo dataset, esto puede ser debido a que las mismas no poseen puntos de ventas o, también, que existan provincias de otros países y dado que el dataset de puntos de ventas sólo considera las que hay en Argentina no habría inconvenientes de que se hayan eliminado estas provincias.

En la quinta unión entre pdv_ubicacion y pais se obtuvo un nuevo dataset denominado pdv_ubicacion. Se eliminaron tres países: Chile (id_Pais 2), Paraguay (id_Pais 5), y Bolivia (id_Pais 7).

Tal como se mencionó anteriormente es de esperar su eliminación dado que no se consideran puntos de ventas de otros países en tal base.

Por último, se unió ventas_producto y pdv_ubicacion se obtuvo un nuevo dataset denominado ventas_producto_pdv. Se observa que existen 53 puntos de ventas que ya no se encuentran en el nuevo dataset.

b. Luego de aplicar el checklist de limpieza, ¿tiene features resultantes de tipo "object"? ¿Qué decisión tomaron al respecto?

Existen tres variables: sku, año y mes que a pesar de ser números deben ser consideradas como string dado que son variables categóricas.

c. ¿Qué decisión se tomó con respecto a los puntos de venta que no tienen información de localización tales como países, provincias y localidades?

Desde el primer informe se había tomado la decisión de eliminarlos. La principal razón para eliminarlas se debió a que representaban un muy bajo porcentaje sobre el total de registros, además suponemos que dado que existen valores faltantes en los mismos registros para las 3 categorías se referían a otro país.

En el caso de conservarlos se podría agrupar en una nueva categoría como "Otros" pero dado que el objetivo final es predecir la demanda mensual en los diferentes países y zonas que opera la compañía esta categoría no nos sería útil. Por último, imputar este tipo de datos puede introducir grandes sesgos dado el tipo de variables que se está analizando.

d. ¿Qué decisión se tomó con aquellas filas cuya cantidad_pedida era menor o igual a 0?

Para aquellas filas cuya **'cantidad_pedida'** era menor o igual a cero quedarán excluidas luego de la detección y eliminación de outliers a través de los percentiles 0.995 y 0.005.

e. ¿El dataset tiene la feature presentación NaN o string vacío? En caso afirmativo, ¿qué decisión consideran pertinente tomar al respecto?

Sí, en el dataset existen una categoría con string vacío y una categoría de "No Asignada".

Para el primer caso dado que los registros son bajos se decidió agruparla con la categoría "No Asignada". En el caso de esta última, no se optó por su eliminación dado que el número de registros era considerable y la eliminación de estos registros nos generaría una pérdida de información en otros campos que sí poseen información. Además, dado el dominio acotado, no estamos en posición de tomar una decisión como esta porque podría estar correcto que existan productos sin algún tipo de marca asignada temporalmente.

f. ¿El dataset tiene la feature marca NaN o string vacío? En caso afirmativo, ¿qué decisión consideran pertinente tomar al respecto?

No se presentan datos nulos o NaN ni string vacíos en el dataset. Sin embargo, se observa una marca denominada "Sin definir" que se presentan en un porcentaje de registros altos como para eliminarla, por lo que se decidió conservarla.

g. Tomaron alguna decisión adicional que redujo la cantidad de filas en el dataset resultante

Sí, en concreto se tomaron tres decisiones. En primer lugar, se realizó un filtro a la base de datos de aquellos productos que eran considerados no comestibles. En segundo lugar, se decidió eliminar aquellos valores de ***'unidadkg'*** por debajo del 0 (o negativos). Por último, se agruparon los pedidos

por mes (el dataset estaba a nivel día), ya que es el horizonte temporal que nos interesa predecir y nos ayuda a reducir bastante el tamaño de la matriz sin perder casi información

5. Enriquecimiento del Dataset sumando nuevas Features

a. Tratamiento de las features presentación, categoría y marca

En este inciso se procede a "acomodar" la información de presentación, categoría y marca. Para ello, se eliminaron los acentos, números, signos de puntuación y se transformó todo a minúscula.

b. One Hot Encoding. Generar variables dummies para aquellas features de tipo objects que considere pueden aportar a los futuros modelos de machine learning. Algunos ejemplos de columnas candidatas a ser columnas dummies son marca, categoría, presentación, provincias-localidad y sku.

En este punto se desarrolla el One Hot Encoding, para su desarrollo se optó por la librería de Sklearn dado que es más eficiente respecto que la opción de Pandas por el tamaño del dataset.

Para el caso de ***'sku'*** es necesario reducir el número de sku al momento de realizar el One Hot Encoding, dado que existen 260 productos y por lo tanto se incrementará mucho el número de columnas . Lo mismo sucede con el feature de ***'ubicacion'***, ya que existen 657 ubicaciones distintas en el dataset por lo que se deberá reducir antes de proceder.

Para la variable ***'sku'*** se decidió mantener el 60% de los productos que representan el total de ventas (quedan 53 sku y una categoría de "otros_sku"), mientras que para la variable ***'ubicacion'*** se seleccionaron las 80 ubicaciones más frecuentes y el resto se agrupó en una nueva categoría denominada "otras_localidades".

c. Valores cantidad_pedida para los periodos t-n lag. Crear columnas nuevas que representen los valores de cantidad_pedida para la combinación de provincia-localidad y sku agrupados por mes para "n" periodos anteriores (lag t-n). Podrá utilizar pandas.Series.shift para lograr este requerimiento. Seleccione el n que considere pertinente y justifique su elección.

En esta sección se crean nuevas columnas para cantidad_pedida para los "n" períodos anteriores. Se definió a "n" igual a 3 para capturar la información del cambio de estación que ocurre cada 3 meses dado que el producto que se está analizando (ver análisis en informe 1) es estacionario, presentando una mayor demanda en las estaciones más cálidas del año. Para ello se agrupó la variable cantidad_pedida según el año, mes, ubicación y sku, mediante el método .sum() y luego se cacularon las nuevas columnas con el metodo shift(). Como al usar este metodo se generan valores NaN para los n primeros meses, estos fueron eliminados del dataset.

d. Crear columnas nuevas para medias móviles simples y/o ponderadas. Crear medias móviles simples y/o ponderadas de la columna cantidad_pedida de n periodos pasados para la combinación de provincia-localidad y sku agrupados por mes. Considere utilizar pandas rolling para el cálculo de medias móviles. Seleccione "window" que considere pertinente y justifique su elección.

En este punto se calculan las medias móviles simples para la columna cantidad_pedida. Cabe recordar que el dataset, tal como se mostró en el informe anterior, presenta una fuerte estacionalidad, en que los meses más cálidos del año aumenta la demanda de los productos y en los meses de temperaturas más bajas la demanda cae. Es importante mencionar esto dado que el

cálculo de las medias móviles se ve afectado por esta estacionalidad e influye directamente en la tendencia de la misma.

Se utilizó una ventana de 2 y 3 periodos anteriores para media móvil, sin ponderación, para capturar la estacionalidad mencionada anteriormente. No se eligieron ventanas mayores para evitar una pérdida innecesaria de información ya que cada paso elimina el último dato de la serie de tiempo, es decir que no es posible obtener datos para Enero y Febrero de 2018.

e. Features externas a agregar

Utilizando información obtenida de Wikipedia, se agregó información de la cantidad de habitantes y la densidad poblacional de cada provincia. Esta variable es de suma importancia para entender de mejor manera y realizar un análisis más preciso de la demanda de productos según la ubicación en que se está realizando, más aún en un país como Argentina en que la densidad poblacional varía en gran cuantía dependiendo de la provincia en la que se encuentre. Ejemplo: las localidades de la provincia de Buenos Aires tendrán una densidad mayor respecto a la provincia de Tierra del Fuego, por lo que hallar una mayor demanda en la primera es esperable.

f. Crear columna totalkg

Al igual que se realizó para el primer informe se crea la columna '***totalkg'*** con el objetivo de unificar la unidad de medida para todos los productos a producir por la fábrica, dado que cada uno de ellos tienen distintas presentaciones. Para ello, se utilizó el producto entre las variables ***cantidad_pedida*** y ***unidadkg***.

Generada la variable, se procede a detectar y eliminar los outliers para esta nueva variable. Para ello se optó por utilizar los percentiles de 0.995 y 0.005 (tal como se había realizado en el punto anterior pero para ***'cantidad_pedida'***). A continuación se observa el boxplot para la variable totalkg sin filtrar outliers y con el posterior filtro de los mismos.

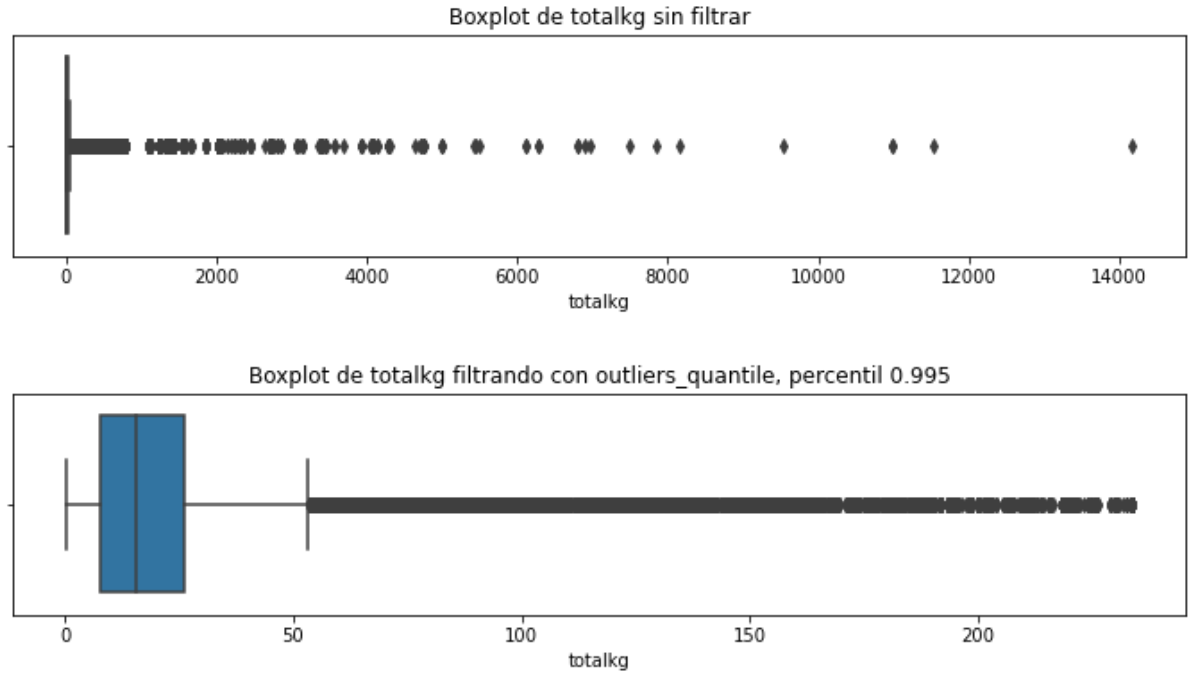

Figura 3 - Boxplot de la variable ***'totalkg'*** con outliers y sin outliers

3 PRINCIPALES CONCLUSIONES

Finalmente, luego de aplicar todos los pasos de la checklist de limpieza y curación de datos, obtenemos un dataset depurado y en un formato apto para utilizar como input en los experimentos de machine learning de los próximos prácticos.

Sin embargo, tenemos en cuenta que este proceso de enriquecimiento es iterativo, ya que cuando empecemos a entrenar modelos para predecir, van a hacerse evidentes nuevas relaciones entre variables o nuevas variables que pueden ser útiles agregar y por lo tanto, probablemente tendremos que rever algunas de las decisiones que tomamos en los primeros dos prácticos de esta mentoría para mejorar las predicciones.