INTRODUCCIÓN A LA CIENCIA DE DATOS

2024

TAREA FINAL

Grupo nº 9
Pedro Ignacio Lasa
Andrés Santos

INTRODUCCIÓN

El presente informe se basa en datos generados en el ámbito de la industria alimenticia, donde uno de los autores del informe se desempeña laboralmente. La gerencia de mantenimiento de la empresa solicitó un análisis respecto a la presunción de un aumento significativo y generalizado en el precio de los repuestos de la maquinaria, desde el inicio de la pandemia. Este supuesto incremento habría afectado notablemente el presupuesto destinado al mantenimiento de la empresa.

Dada la naturaleza sensible de los datos, nos reservamos el nombre de la empresa de donde se obtuvieron, así como la de los proveedores asociados a las compras de los repuestos.

Se comenta brevemente el tratamiento en los datos realizados, los problemas encontrados resueltos, y los problemas que no pudieron ser resueltos. Además, se presentan algunos resultados obtenidos y otros análisis, que podrían realizarse aplicando las técnicas aprendidas en el curso.

BASE DE DATOS

La empresa en cuestión cuenta con un sistema de gestión robusto y flexible, hace posible obtener varios datos sobre las compras de suministros que se realizan. Dada la abundancia de los datos, fue importante determinar qué información era realmente relevante para realizar nuestro análisis.

Se definió hacer una búsqueda de “compras por material” con las siguientes variantes de búsqueda:

- Fecha de documento de compra: entre el 01/01/2020 a 30/06/2024
- Centro (o planta): todas las plantas que conforman la empresa.
- Nº de material: se indicó el rango de los números de material reservado para los repuestos.

Una vez ejecutada la transacción correspondiente, se obtienen los datos requeridos en un data frame de 80.472 instancias y 15 columnas, con las siguientes variables:

Nombre Columna	Formato	Descripción
Documento compras	numeric	Nº que identifica la compra. (Cada compra puede tener varias posiciones, una para cada material).
Fecha documento	POSIXct	Fecha en que se crea el “Documento de compra”.
Estado liberación	character	Se indica con “X” si la compra fue liberada, o en caso contrario está vacío.
Material	numeric	Nº único que identifica al material. Este número es designado por la empresa cuando se adquiere el material por primera vez y se da de alta en el maestro de materiales de la empresa.

Texto breve	character	El texto breve da una descripción breve del material. Esta descripción también se genera al dar de alta el material en el sistema de gestión y debe cumplir con ciertas reglas.
Cantidad de pedido	numeric	Cantidad de unidades pedidas, tomando como unidad la “unidad de medida pedido”.
Unidad medida pedido	character	Unidad de medida del pedido, p. ej.: “C/U” (unidad), “M” (metro), “L” (litro), etc.
Centro	character	Indica el centro o planta para el cual se realiza el pedido de compra. Esta información también puede ser sensible.
Valor neto de pedido	numeric	Indica el valor neto de la posición del pedido del material asociado. Si se compran 5 unidades del material “X”, será el precio por las 5 unidades.
Moneda	character	Moneda con la cual se realiza la compra y con la cual se expresa el “Valor neto pedido” y el “precio neto”. Las monedas en este caso pueden ser UYU (pesos uruguayos), USD (dólares americanos), o EUR (euros).
Precio neto	numeric	Precio de la unidad de pedido expresado en la moneda del campo “Moneda”
Proveedor	character	Nº de proveedor y nombre del proveedor información altamente sensible.
Cantidad base	numeric	Cantidad de medida base.
Grupo de artículos	character	Código que indica el tipo de material instrumentación, mecánico, eléctrico

TRATAMIENTO DE DATOS

Nuestro data frame con las variantes mencionadas, se exportó a un archivo Excel y luego se realizó el siguiente tratamiento de los datos utilizando el software R:

- **Eliminación de líneas de compras con indicador de borrado y sin liberación de compra:**

  Se eliminaron las líneas de compras marcadas con el indicador de borrado, ya que son gestiones que por alguna razón quedaron truncas. Además se eliminan las instancias sin marca en el “estado de liberación”, ya que son compras que nunca llegaron a emitirse.

- **Conversión de Tipo de Cambio:**

  Fue necesario pasar toas las compras a una moneda común. Se definió usar los dólares americanos como moneda común. Para ello se agregó una columna en nuestro data frame con el tipo de cambio correspondiente a la “Fecha de documento”. Se extrajo un data frame de yahoo con los tipos de cambio EUR/USD y UYU/USD, utilizando la fórmula:

  getSymbols("EURUSD=X", src = "yahoo", from = "2020-01-01", to = Sys.Date())

  getSymbols("UYUUSD=X", src = "yahoo", from = "2020-01-01", to = Sys.Date())

  Se agregó una columna con el tipo de cambio según la compra haya sido en euros o en pesos uruguayos.

Para hacer el “merge” y hacer coincidir el tipo de cambio con la “fecha del documento” fue necesario armonizar el formato de la fecha en ambos data frames.

- **Cálculo del Precio Unitario en dólares americanos:**

  Se añadió una columna con el precio unitario en dólares americanos, utilizando la siguiente fórmula:

  $$Precio\ unitario\ USD\ = \frac{(Valor\ neto\ de\ pedido)\times(Tipo\ de\ cambio)}{Cantidad\ de\ pedido}$$

- **Clasificación de las Compras:**

  Las compras se clasificaron según se hayan realizado antes o después de la fecha 01/01/2022, la cual se tomará como fecha de corte (FC).

- **Cálculo del Precio Unitario Promedio:**

  Para cada material, se calculó el precio unitario promedio antes y después de la fecha de corte, y se determinó el porcentaje de variación utilizando la siguiente fórmula:

  $$Cambio\ porcentual\ = (\frac{precio\ promedio\ despues\ de\ FC}{Precio\ promedio\ antes\ de\ FC} - 1) \times 100$$

- **Generación de Nuevo Data Frame**

  Se creó una nueva base con los campos: número de material, precio promedio antes FC, precio promedio después FC, y cambio porcentual. Nuestro nuevo data frame , que llamamos “comparación_precios”, tiene 7493 instancias y 4 variables. Aquí, nos quedamos solo con las líneas de repuestos que tienen compras antes y después de la FC, ya que si no cuenta con alguno de los precios, no hay comparación posible.

DATOS ATÍPICOS

Los datos atípicos o “outliers” distorsionan la muestra de datos; en el nuevo data frame, “comparación_precios”, llama la atención algunos casos en que compras asociadas a los mismos materiales cuentan con diferencias de precio unitario del orden del 300%, y que inclusive pueden ir hasta el 1000%. Aquí algunos ejemplos:

Estas diferencias de precios, un tanto inconsistentes con la realidad, se pueden explicar por el hecho de haber utilizado el mismo número de material para gestionar la compra de materiales diferentes. Se puede tratar de materiales con descripciones ambiguas o vagas en el “texto breve”, que lleven a ser utilizadas para la compra de materiales muy diversos. Se estudiaron algunos ejemplos puntuales para poder corroborar esto, y se confirma dicha hipótesis.

Lo descrito anteriormente, es un claro problema que se asocia a la calidad en los datos. Limpiar nuestro data frame de estos errores implicaría un trabajo extenso y tedioso pero necesario. La opción que se tomó en este caso, un tanto rudimentaria, fue la de filtrar los materiales cuya variación de precio unitario no supere un determinado margen que se considera razonable.

ALGUNOS RESULTADOS OBTENIDOS

En base al tratamiento de datos anterior se hicieron histogramas (que no se presentarán aquí), y se calculó el aumento de precios promedio, así como el aumento promedio ponderado por la cantidad de repuestos consumidos, llegando a concluir que tal aumento generalizado sí existió y fue del orden del 20%. No se presentan mayores resultados considerándolos sensibles e irrelevantes para los objetivos de la presente tarea.

POSIBLES MODELOS A APLICAR:

Si el investigador desea puede implementar posibles metodologías de cluster o modelos que intenten predecir el precio de un artículo, como también las variables que mayormente impactan en el mismo.

1) La primera técnica posible es el **Análisis de Componentes principales (PCA),** técnica de reducción de dimensionalidad que transforma un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas. Es posible reducir la complejidad del conjunto de datos y encontrar patrones subyacentes. El PCA permite identificar las variables más significativas que logren explicar la mayor parte de la varianza en el conjunto de datos. **Variables a utilizar:** Se pueden utilizar todas las variables numéricas.

2) Agrupamiento de datos por **K-Means**, técnica de agrupamiento que asigna cada punto de datos a uno de los "k" grupos predefinidos. Tiene como objetivo identificar grupos homogéneos de compras o materiales en base a sus características, segmenta el conjunto de datos en grupos con características similares. **Variables a utilizar:** se pueden utilizar variables numéricas y categóricas, como "Cantidad de pedido", "Valor neto de pedido", "Grupo de artículos", "Centro".

3) **Modelo Logit:** el modelo Logit predice la probabilidad de un evento binario (por ejemplo, “aumento de precio" o "baja de precio"). **Variables a utilizar:** se pueden utilizar variables numéricas y categóricas, como "Valor neto de pedido", "Cantidad de pedido", "Grupo de artículos", "Centro", "Proveedor” (las variables se suelen recopilar como 1 o 0).

4) **Regresión Lineal Múltiple:** el modelo de regresión múltiple predice una variable continua (por ejemplo, "valor neto de pedido") en función de variables predictoras o independientes. Además, permite identificar los factores que influyen en el valor neto de pedido de una compra (el investigador puede interpretar el impacto de cada variable independiente en la variable dependiente). **Variables a utilizar:** Se pueden utilizar variables numéricas y categóricas, como "Cantidad de pedido", "Precio neto", "Grupo de artículos", "Centro", "Proveedor".