Proyecto Walmart

Beto, Cesar, Luis y Paola

2019-12-18

Contents


```
#install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
library("bookdown")
```

Chapter 1

Breve descripción del proyecto

El presente trabajo busca explorar un conjunto de datos de los clientes de Walmart, a través de técnicas de análisis de datos y aprendizaje de máquina, con el objeto de agrupar los tipos de visitas que hacen los clientes en torno a una serie de categorías desarrollas por la empresa de manera interna, empleando la metodología CRIPS-DM.

Chapter 2

Comprensión del negocio

2.1 Antecedentes

Uno de los mayores intereses de las cadenas comerciales es conocer el comportamiento de sus clientes con el objeto de poder implementar estrategias de venta acordes a los diferentes necesidades existentes en el público.

En este sentido, un enfoque que puede ser de utilidad es segmentar las visitas que los clientes efectúan en los establecimientos en diferentes tipos de viajes; ello significa determinar un conjunto categorías que reflejen los motivos que se encuentran detrás de una visita a un establecimiento. Para tal efecto, se puede echar mano de la información histórica que se tenga sobre el cliente, en términos de 1) los datos personales de este (por ejemplo, su información socio-económica) y 2) las transacciones de artículos que este haya realizado (tanto para adquirir artículos como para devolverlos por algún motivo); de manera que sea posible delinear un patrón de comportamiento que permita extraer elementos de valor para personalizar experiencias de comprar acordes a la realidad del mercado.

A manera de ejemplo, se puede pensar que existen diferencias entre los clientes que hacen una visita rápida a un establecimiento para comprar dulces, que aquellos que surten sus alacenas para consumo de víveres en la de semana.

Como parte de este interés, Walmart, la cual es una cadena comercial con amplia presencia alrededor del mundo en la venta de artículos de tecnología, hogar, linea blanca, supermercado y muchos otros, ha desarrollado una metodología a nivel interno que le permite agrupar en torno a 38 categorías a las diferentes visitas que realizan sus clientes, en función de los artículos que los clientes adquieren.

En tal contexto, durante 2015, esta cadena hizo disponible, a través de Kaggle, un conjunto de datos de las visitas de sus clientes[1], que reflejan dicha categorización de los visitas de sus clientes, con el objeto de incentivar a científicos de

[1] Véase https://github.com/Kaggle/kaggle-api

datos a recrearla y explorar si ésta puede refinarse, redundando en un proceso de mejora en la segmentación de su clientela.

Es así que la idea de este proyecto es analizar los datos aportados por Walmart para proponer un enfoque que permita recrear la categorización hecha por esta empresa, basándose en métodos de aprendizaje de máquina a través de la metodología CRIPS-DM, mediante los conceptos vistos en el curso de Minería y Análisis de Datos.

2.2 Determinación del objetivo

El objetivo de este proyecto proponer una metodología para resolver el problema de clasificación de los clientes de Walmart, presentes en las bases de datos que dicha empresa compartió a Kaggle en el año 2015, a partir del análisis de datos y métodos de aprendizaje de máquina.

2.3 Determinación de criterio de éxito

Como criterio para determinar el cumplimiento del objetivo del proyecto, se estableció que el modelo propuesto tenga un mejor desempeño que el bechmark de la referida competencia de Kaggle.

2.4 Plan del proyecto

En línea con la exposición previa, a continuación se presenta el plan de proyecto para lograr el objetivo de este proyecto, mismo que se llevará a cabo a través de las fases que se describen, en alto nivel, a continuación:

- **Comprensión de los datos de Walmart:** conformada por las etapas de 1) extracción de la información aportada por Walmart, en su estado puro (es decir, datos crudos, sin realizar ningún tratamiento de la información); y 2) el estudio de las variables que componen el conjunto de datos de Walmart, junto con un proceso de exploración de la información contenida en ellos, en términos de una variable, pares de variables o múltiples combinaciones de ellas.
- **Preparación de los datos:** consistente en el proceso de selección e integración de los datos que serán útiles para plantear un metodología de clasificación de los clientes de Walmart, así limpieza de los datos (consideran la imputación en el caso de variables con información no disponible o ausente por algún motivo), y la ingeniería de características pertinente para mejorar el desempeño del enfoque propuesto.
- **Modelado:** corresponde al diseño de un conjunto de modelos, basados en aprendizaje de máquina, encaminados a resolver el problema de clasificación de los viajes de los clientes de Walmart, así como los criterios con-

siderados para seleccionar el modelo con el mejor desempeño y el ajuste realizado para calibrar sus hiper-parámetros.

- **Evaluación:** se refiere a la etapa en donde se realiza la evaluación del modelo óptimo seleccionado como óptimo en la etapa previa, su posterior re-entrenamiento tras conjuntar los datos de entrenamiento y prueba con hiper-parámetros optimizados así como el reporte de la posición final en el tablero de Kaggle del desempeño logrado.
- **Implantación:** relativo al desarrollo de web service en flask para predecir resultados con el modelo final a partir de nuevos datos, así como el reporte ejecutivo que relata los principales hallazgos e hitos del proyecto.

Para mejor referencia, se provee un repositorio en Github https://github.com/paola-md/Walmart-Data-Mining- que conjunta los archivos de trabajo realizados con motivo del proyecto en cuestión, particularmente la totalidad de scripts en Bash, R y Python, así como las instrucciones a través de las cuales se podrá replicar el contenido del proyecto.

Chapter 3

Comprensión de los datos

3.1 Recolección de datos

De acuerdo a la documentación disponible[1], uno de los requisitos necesarios para descargar los datos de visitas de los clientes de Walmart, es aportar la credenciales de un usuario registrado en el Kaggle.

En este caso, para facilitar este proceso, se implementó un programa en Bash, denominado download_extract_data.sh, el cual aprovecha la herramienta Wget de UNIX, para realizar la descarga de los datos del sitio electrónico en comento, recibiendo un archivo de configuración con las credenciales de autenticación de un cierto usuario (archivo cookies.txt).

A su vez, se creó un script en R (wrinting_feather.R, presente en la carpeta /data) que convierte los archivos .csv a formato .feather.

Como resultado, dentro de la carpeta /data se obtiene los siguientes archivo en formato .csv y .feather:

- **test.csv:** el cual contiene los datos del conjunto de entrenamiento,
- **train.csv:** relativo a los datos de entrenamiento, y finalmente,
- **sample_submission.csv:** es un ejemplo del formato en que se deben aportar los datos al sistema de Kaggle para la evaluación del desempeño del modelo de clasificación propuesto para el problema que nos ocupa.
- **test.feather**,
- **train.feather**.

** Instrucciones para descarga de datos**

1. Debemos darle permisos de ejecución al script de Bash que se encuentra en la carpeta */build/comprension_datos* desde la terminal:

[1] Véase https://github.com/Kaggle/kaggle-api

```
chmod +x download_extract_data.sh
```

2. Posteriormente ejecutamos dicho programa:

```
./download_extract_data.sh
```

3. Como se ha mencionado, el resultado de la ejecución de este script es la descarga de tres archivos dentro de la carpeta /data (train.csv, test.csv y sample_submission.csv).

3.2 Análisis exploratorio de datos

De acuerdo a la documentación de Kaggle, la información de Walmart se proporciona en términos de las siguientes 7 variables:

- **TripType:** Variable objetivo. Son 38 diferentes categorias,
- **Visit Number:** código identificardor de la visita de un usuario al establecimiento (o simplemente, "viaje"),
- **Weekday:** Día de la semana en que ocurrió el viaje del cliente,
- **UPC** corresponde al número de barras de cada producto (es decir, un código identidicador del mismo),
- **ScanCount:** relativo al número de productos involucrados en la transacción del cliente. Cabe descatar que si se trada de una devolución, se representa como un número negativo
- **Department Description:** corresponde a una descripción de la categoria a la que pertenece el producto involucrado en la transacción del cliente.
- **Fineline Number:** relativo a un código identificador de 5,196 productos.

3.2.1 Univariado

3.2.2 Bivariado

3.2.3 Multivariado

Chapter 4

Preparación de los datos

Este documento describe el proceso de preparación de los datos llevado a cabo sobre la información de Walmart. Cabe destacar que este comprende las etapas:

- **Selección e integración de datos**,
- **Limpieza de datos**,
- **Ingeniería de características**.

A continuación se describirá a mayor detalle cada uno de los puntos referidos. El detalle de la implementación de tales procesos se puede ver a través del archivo **DataPreparation.R**.

4.1 1. Selección e integración de datos

Para esta etapa se debe destacar que, como fue expuesto en la sección del análisis exploratorio de datos, la base de datos de Walmart provee información de los viajes de los clientes en una manera desagregada conforme a las visitas de tales individuos a las tiendas, en función de cada uno de los artículos que se compraron en una visita. Es decir, para cada visita de un cliente pueden existir múltiples reglones, los cuales refieren a los artículos que se adquirieron o devolvieron en dicho evento.

Desde el punto de vista del funcionamiento de los modelos de aprendizaje de máquina, esto constituye una limitante puesto que, en general, tales realizan tareas bajo la idea de que las unidades observacionales en estudio, en este caso las visitas de los clientes a la tienda, aparecen de manera única en las tablas que guardan la información correspondiente.

Es así que la primera decisión fue transformar la información a una tabla en donde se agruparan las visitas de los clientes (ver sección de ingeniería de características para mayor detalle).

4.2 2. Limpieza de datos

En complemento se llevaron a cabo las siguientes acciones de limpieza sobre los datos aportados por Walmart:

- Sobre los campos que contienen información tipo texto, se aplicó una transformación que convierte todo a minúsculas.
- Se realizó la imputación de valores cero sobre los campos donde existe información sobre las variables *Upc* y *FinelineNumber*.

4.3 3. Ingeniería de características

Sobre la base de datos en comento se creó un conjunto de nuevas variables con el propósito de nutrir con mayor elementos de información al modelo que se busca implementar. Entre tales, se encuentran:

- En primera, se identificó cuales fueron los artículos que se involucraron en la transacción de un cliente (usando el identificador de la misma),
- Posteriormente, se calculo la proporción de cada articulo que representan respecto al volumen de aquellos involucrados en la transacción (tanto de aquellos adquiridos como devueltos),
- Usando la información anterior, se consolida una nueva base de datos con la información de cada visita de un cliente a través de un renglón en donde se indica que proporción tuvo cada uno de los artículos que se vieron involucrados el evento.

Adicionalmente, se construyeron nuevas características:

- A través de una nueva variable (*obj_abs*) que refleja la cantidad total de objetos que se involucraron en la transacción, donde aquellos que se adquirieron se representaron de manera agregada con signo positivo y los que se devolvieron con signo negativo.
- La variable *num_obj* refleja el valor absoluto de la cantidad total de objetos que se involucraron en la transacción,
- Se creó una variable indicadora(*prod_miss*) que nos dice si en la base original de Walmart se encontraba ausente el campo *Upc*.
- Además, no se consideraron las columnas *Weekday*, *Upc* y *visitnumber*

Por otra parte, también se llevaron a cabo las siguientes acciones:

- Se creó una variable (*weekend*) que refleja si la transacción de un cliente se llevó a cabo o no en un fin de semana,
- En adición, se generó una variable (*day*) que codifica a manera de categorías numéricas el día de la semana (*Monday*=1, *Tuesday*=2, ..., *Sunday*=7),
- Análogamente, se construyó una variable numérica (*departmentdescription*) que codifica con categorías numéricas a los diferentes departamentos a los cuales pertenencen los artículos de la tienda.

- En lo tocante a la devolución de artículos, se crearon dos nuevas variables a) *devol* que indica si en la transacción existió un evento de devolución de artículos, y b) *porc_devol* que refleja la proporción de artículos devueltos con respecto a aquellos que se involucraron en la transacción de un cliente,