PRA2: Limpieza y análisis de datos

M2.851 - Tipología y ciclo de vida de los datos

Víctor Blanes Martín
Carlos Allo Latorre

28/05/2021

Tabla de contenido

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Hoy en día, la compra de dispositivos electrónicos y en especial de portátiles, está a la orden del día y es más frecuente que nunca. A raíz del desarrollo de nuevas tecnologías, cada vez es más frecuente disponer de una oferta de productos más amplia, excesiva en ocasiones, que dificulta la toma de decisiones en cuanto a la compra de dichos productos.

En función del perfil del comprador, el desconocimiento de la gama de productos y sus características puede que los lleve a tomar una decisión de compra poco adecuada o ajustada en precio. Por este motivo, sería interesante conocer qué variables o características son más influyentes en el precio a la hora de compra de un ordenador, para que de esta manera, el comprador:

- Pueda hacerse una idea del presupuesto aproximado que tendrá el ordenador que desea en base a las características técnicas deseadas.
- Pueda conocer si es verdad que algunas marcas, como Apple, tienen un precio algo superior al resto de las marcas.
- Pueda conocer qué características son las más influyentes en el precio para en base a estas, poder centrarse en lo que realmente necesita para incrementar o decrementar su presupuesto.

Para ello, se podría usar el dataset de ordenadores portátiles construido en la práctica 1, en donde, entre otros objetivos a responder, el recién presentado era uno de ellos. Sin embargo, tras un comienzo con este dataset, surgió el problema de que no poseían suficientes datos ni características como para realizar buenos análisis ni modelos, por lo que se decidió buscar bajo el mismo objetivo a tratar, otro dataset. Tras la búsqueda se seleccionó el conjunto de datos "Laptop Prices" de Kaggle (https://www.kaggle.com/ionaskel/laptop-prices), que posee 1303 registros de ordenadores con sus correspondientes características y precio.

2. Integración y selección de los datos de interés a analizar.

La integración de los datos será mínima ya que todos los datos a usar provienen de un mismo dataset. Como se usará el lenguaje de programación Python, en términos de programación la única importación que se realizará será la de carga del dataset al entorno, que se realizará mediante pandas apuntando al archivo comentado, que ha sido importado al GitHub del proyecto. Cabe resaltar, que si se deseara usar dos datasets diferentes con las mismas características, se debería de realizar una integración de estos, donde habría que tener en cuenta aspectos como posibles repeticiones de objetos, que las propiedades se presenten en las mismas unidades... Este proceso se realizó en la práctica 1, en el momento en el que se integraban dos datasets diferentes (uno de cada web en donde se realizó WebScraping), en uno sólo.

Respecto a la selección de los datos, nos quedaremos con todas las filas, ya que cada una de ellas corresponde a un ordenador diferente y aporta información al estudio que se está realizando. Además, nos encontramos ante un número de ordenadores (aproximadamente 1000) no tan grande como para tener que hacer reducción de la cantidad. Sin embargo, para el caso de tener que aplicar dicha técnica, consideramos que las dos mejores formas de hacerlo serían el método de muestra aleatoria simple sin sustitución (para no tener repeticiones de ordenadores en el dataset resultante), o muestra de clústeres, donde cada clúster podría estar correspondido por la marca o por intervalos de precio para asegurar que tenemos muestras de todos los precios.

Sobre las columnas, encontramos que la primera de ellas que no nos da ninguna información útil para el estudio y ser un simple id ascendente, con lo cual la eliminaremos. Haremos lo mismo con la columna que se refiere al modelo del producto, pues el objetivo en todo momento es realizar una comparación en base a características o marcas, pero no en base al modelo del portátil directamente.

Por tanto, tras esta limpieza, las columnas que resultarán del dataset junto con su significado según se proporciona en el repositorio original serán:

- Company --> Company Name
- TypeName --> Laptop Type
- Inches --> Screen Inches
- ScreenResolution --> Screen Resolution
- Cpu --> CPU Model
- Ram --> RAM Characteristics
- Memory --> Memory
- Gpu --> GPU Characteristics
- OpSys --> Operating System
- Weight --> Laptop's Weight
- Price_euros --> Laptop's Price

3. Limpieza de los datos.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? Tras la carga inicial, se aprecia como no se encuentra ningún valor vacío o cero. Sin embargo, tras el proceso de limpieza que se comentará posteriormente, se encuentran valores nulos en la variable “ScreenResolution_Type”, que será creada a partir de la columna ScreenResolution, al no presentarse el tipo explícitamente en la variable inicial.

Para tratar estos nulos, se sustituirán los valores perdidos por una misma constante o etiqueta, en este caso, "Unknown". Se realiza de tal forma ya que falta un gran número de registros (un 33%), y el uso de otras técnicas como la sustitución por la moda o mediana harían que tuviéramos muchísimos datos 'no reales'.

Asimismo, tampoco tiene sentido aplicar técnicas de sustitución basadas en modelos de predicción, ya que la resolución de la pantalla no es una característica que dependa del resto de variables de la muestra.

Con respecto al resto de variables, tras realizar un pequeño estudio, se aprecia que no hay datos perdidos o que indiquen la pérdida de valor. Si bien es verdad que para sistema operativo encontramos 'No OS', que puede dar lugar a confusión. Sin embargo, este valor es válido, ya que algunos ordenadores pueden no tener sistema operativo preinstalado y, posiblemente, esto sea algo que abarate el coste del mismo.

3.2 Preparación de datos.

Nos hemos sentido libres de añadir este apartado, por el hecho de que antes de pasar al apartado 4 en donde se analizan más profundamente los datos y se elaboran modelos o incluso del apartado 3.3 consistente en la búsqueda de valores extremos, se ha considera que es necesario un proceso de preprocesado o data cleaning de los mismos.

Para ello, se han centrado los esfuerzos en variables como Ram o Weight, que son variables que se han de tratar como numéricas pero que al tener la unidad de medida en cada uno de sus valores, dificulta su tratamiento. Para ello, se ha comprobado que todos los valores estaban indicados en las mismas unidades, GB y Kg respectivamente, y se ha dejado únicamente el valor numérico como contenido del campo. Además, se ha modificado el nombre de las columnas para indicar las unidades Ram(GB) y Weight(Kg).

Siguiendo en la misma línea, se ha realizado un análisis del contenido de la variable memoria RAM. Se aprecian memorias individuales y también compuestas (híbridas) por varios tamaños (GB y TB) y tipos de tecnologías (SSD, Flash Storage, HDD e Hybrid). Para tratar este hecho, se han transformado todas las medidas a GB y se han creado cuatro nuevas columnas cuyo contenido será numérico: MemorySSD(GB), MemoryHDD(GB), MemoryFlash(GB), MemoryHybrid(GB).

Para el caso de la CPU, se ha identificado que el patrón que sigue esta columna es el siguiente: en primer lugar se da el fabricante, en segundo lugar la versión que se proporciona y por último, la velocidad de reloj de la misma. Por tanto, realizando este una separación de estos 3 campos, resultan las columnas CPU_Company, CPU_Version, CPU_Speed(GHz).

En la misma línea con la GPU, resulta claro ver cómo esta presenta la estructura de fabricante de la GPU seguido por la versión, por lo que da lugar a las nuevas columnas GPU_Company, GPU_Version.

Finalmente, con la resolución de pantalla se observa el siguiente patrón: en primer lugar, en algunas ocasiones, se describe cualitativamente la resolución de la pantalla (p.ej.: Full-HD), seguido por el tamaño de pantalla en píxeles con el patrón alto x ancho. Tal y como se ha realizado anteriormente, se han separado estos valores en 3 columnas: ScreenResolution_Type, ScreenResolution_Width, ScreenResolution_High.

Una vez se ha llegado a este punto, ya se poseen datos más preparados para poder realizar análisis posteriores, y muchas variables numéricas listas para trabajar con ellas. De forma previa a realizar este procesamiento, muchas de las variables estaban en formato string ya que incluían la unidad de medida o información adicional en el propio valor. Esto nos impedía su uso para hacer buenos análisis o modelos al ser muchas de ellas variables categóricas y no numéricas.

3.3 Identificación y tratamiento de valores extremos.

Se analiza en primer lugar la variable precio, variable numérica más importante de nuestro estudio. Se aprecia como aparentemente, aunque el percentil 50 está en 977 y su 75 en 1487, la media es de 1123, lo que es indicio de que, tras el percentil 75 encontraremos algún valor más elevado, hasta llegar al máximo de 6099.

Gracias a las representaciones realizadas en el Notebook, se aprecia como hay una gran concentración de datos en la zona en torno a 1000 euros, y que va disminuyendo hasta llegar hasta los 3000, donde los precios empiezan a encontrarse más distantes entre sí hasta llegar a los 6000, dato que podríamos plantearnos como punto alejado o outlier.

Sin embargo, al analizar estos puntos 'alejados' para estudiar si realmente son outliers que hay que eliminar o si son datos posibles, se aprecia que la gran mayoría de portátiles de precios elevados están catalogados como ordenadores Gaming. Este tipo de ordenadores destacan por necesitar de una potencia gráfica y de procesamiento muy superior a la media, hecho que también hace incrementar su precio.

Tras una rápida consulta de precios en el mercado de ordenadores Gaming, confirmamos que no es descabellado que se den dichos precios en este sector específico, con lo que hemos considerado que han de mantenerse en el dataset.

Además, realizando el mismo estudio para el caso del peso se aprecia que ocurre algo similar, donde algunos ordenadores se alejan bastante de la mayoría por la parte superior. Si analizamos la tipología de son estos ordenadores, corresponden también a ordenadores Gaming, conocidos por unas pantallas más grandes por lo general y que requieren de un cuerpo mayor para disponer de una mayor capacidad de disipación del calor, hecho que también hace aumentar su peso. En base a estos análisis, no hemos considerado necesario considerar ningún valor del peso como outlier.

4. Análisis de los datos.

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

A lo largo de este apartado y los siguientes, se han realizado análisis tanto cuantitativos como cualitativos que permiten ver qué variables son las mejores candidatas para formar parte de un modelo o estudio que permita responder a las dudas/objetivos que habíamos planteado al inicio, es decir, que el comprador:

- Pueda hacerse una idea del presupuesto aproximado que tendrá el ordenador que desea en base a las características técnicas deseadas.
- Pueda conocer si es verdad que algunas marcas, como Apple, tienen un precio algo superior al resto de las marcas.
- Pueda conocer qué características son las más influyentes en el precio para en base a estas, poder centrarse en lo que realmente necesita para incrementar o decrementar su presupuesto.

En este sentido, se ha evaluado la correlación de las variables numéricas respecto a la variable dependiente del precio, para identificar cuales de ellas tienen más influencia en la variabilidad del precio.

Respecto a las variables categóricas, se han escogido varias de ellas para representar un boxplot que muestra visualmente, en función de los valores de las mismas, en qué rangos de precios se sitúa cada una. En base a los resultados de los análisis visuales, se han escogido ciertas variables categóricas para cargar al modelo que parecen mostrar una variación más grande en precios en función del valor que tienen.

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para poder responder al segundo objetivo de la práctica, saber si los productos de Apple son más caros de media que el resto de fabricantes, se ha optado por llevar a cabo un contraste de muestras. Para ello, era necesario estudiar la normalidad y homogeneidad de la varianza de la variable del precio en los subsets de productos de Apple y del resto.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.
En base a los resultados obtenidos en las gráficas de normalidad, no se ha podido confirmar que la distribución de estas se corresponda con una normal, con lo que se ha tenido que estudiar la homocedasticidad y el contraste de muestras con pruebas no paramétricas (Fligner-Killeen y Mann-Whitney respectivamente).

El resultado del contraste de muestras nos indica que podemos afirmar que los productos de Apple tienen un precio, de media, superior al resto de dispositivos.

Posteriormente, en base a las variables normalizadas (tanto numéricas como categóricas) que se han identificado tenían más correlación con el precio, se ha entrenado un modelo de regresión lineal que trata de responder al objetivo 1 de la práctica. No obstante, los primeros resultados de esta regresión no arrojan un resultado demasiado positivo respecto a la precisión del modelo.

De cara a la entrega final de la práctica, se estudiará la aplicación de otros posibles modelos que puedan llegar a obtener un mejor rendimiento en este aspecto.

5. Representación de los resultados a partir de tablas y gráficas.
DUDA: ¿Qué se espera exactamente de este apartado? En el Notebook que se ha desarrollado la práctica se han elaborado varias gráficas e incluso comentado y analizada las mismas. ¿Este apartado consiste únicamente en graficar los rendimientos del modelo?

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
Si seguimos un orden basado en los objetivos que nos hemos planteado, para el primer objetivo podemos afirmar lo siguiente:

- Se ha obtenido un modelo que es capaz de aproximar, en la mayoría de ocasiones, un precio justo para el tipo de producto que se está tratando. Aún así, hay desviaciones en varias ocasiones respecto al precio esperado, dado que el modelo no está arrojando unos buenos valores de rendimiento. De cara a la entrega final se estudiará la posible aplicación de otros modelos predictivos.
- Se ha podido confirmar, en base al contraste de muestras realizado, que los dispositivos de Apple son de media más caros que el resto de dispositivos de otros fabricantes.
- Se ha podido verificar cuales de las características de los portátiles eran más influyentes a la hora de subir o bajar el precio del producto. La característica más correlada con el precio es la de la memoria RAM.

Asimismo, se ha podido comprobar gracias al análisis de outliers que la categoría de portátiles más caros es la de Gaming, así como la que tiene componentes de mejor prestación pero que, a su vez, tiene los portátiles más pesados.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.
Se puede encontrar el código desarrollado en el siguiente enlace de GitHub, dentro de la carpeta src. Aquí, se encuentra el Notebook desarrollado en formato .ipynb, al igual que un archivo con extensión .html para facilitar su corrección.

https://github.com/carlosalloUOC/PRA2-Limpieza-Analisis

8. Contribuciones

Contribuciones	Firma
Investigación previa teórica (lectura material UOC, tutoriales y documentación de limpieza de datos, revisión ejemplos anteriores, etc.)	VB, CA
Investigación y elección dataset (estudio de sus características e idoneidad para llevar a cabo los objetivos).	VB, CA
Desarrollo limpieza de datos	VB, CA
Desarrollo análisis de los datos	VB, CA
Elaboración informe	VB, CA