UNIVERSIDAD
TORCUATO DI TELLA

Informe Taller 4

Introducción a la Data Science

Universidad Torcuato Di Tella

Prof. Ramiro H. Gálvez

Hayduk, Gaspar - 18D507
Gonzalez Lelong, Luis - 18Z165

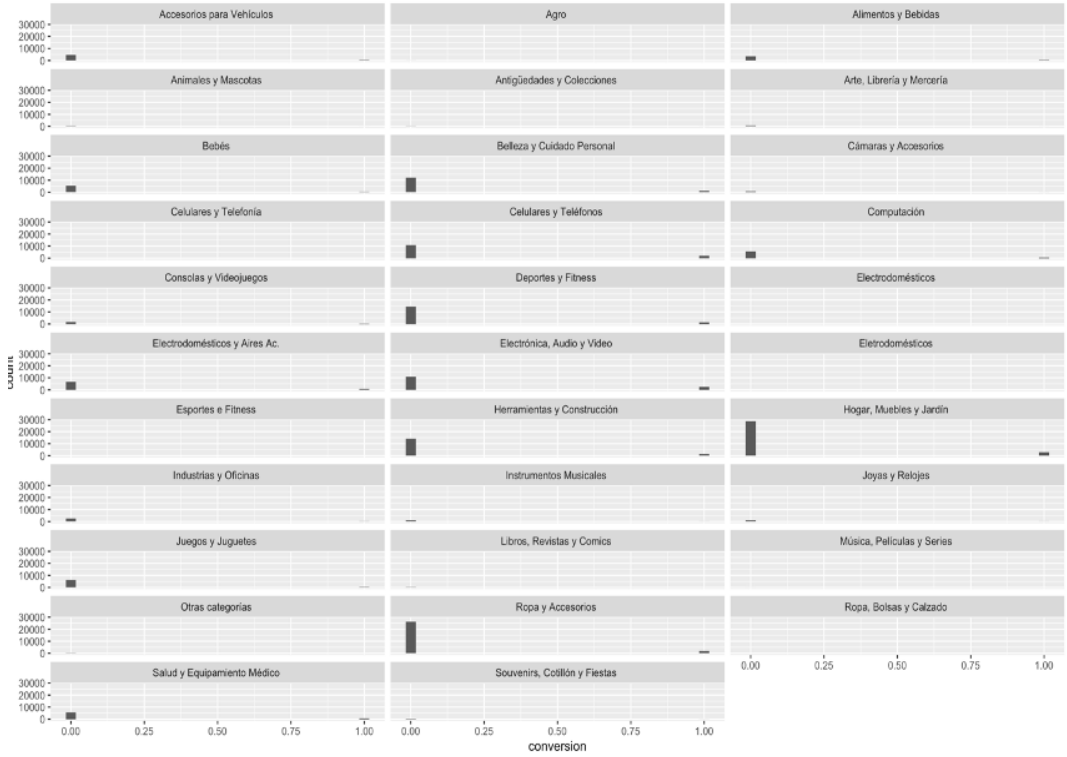
2022

Análisis Exploratorio de los datos.

Para la parte de los gráficos, nos concentraremos en las variables predictoras que sean factores, porque son a partir de este tipo de variables que vamos a hacer *One-Hot-Encoding* y *Bag of Words*.

En primer lugar, nos pareció interesante explorar la categoría del producto (variable "*full_name*"). Viendo que una categoría tiene varias subcategorías, solo nos quedamos con la principal como variable predictora. La hipótesis detrás de esto es que hay ciertos productos que se venden más que otros.

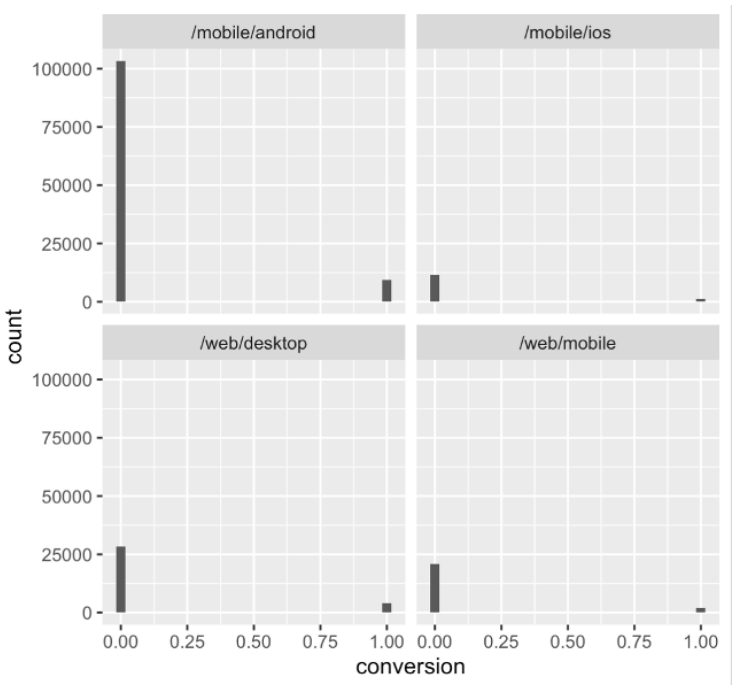
Una vez que consideramos la categoría principal y pasamos la misma a factor, podemos realizar un histograma de la variable "*conversion*" (nuestra variable a predecir) por categoría. El resultado es el siguiente:



Podemos ver cómo la distribución de "*conversion*" es diferente para cada categoría. Más específicamente, la siguiente tabla muestra la proporción de productos para los cuales hubo conversión para las diferentes categorías (la columna 1 indica conversión). En la misma podemos ver que las categorías "Electrónica, Audio y Video", "Celulares y Teléfonos" y "Antigüedades y Colección" son las que tienen mayor proporción de productos con compras.

En segundo lugar, otra variables interesante es “*platform”,* la cual indica desde qué plataforma se está conectando el usuario. Pasando la misma a factor podemos realizar un histograma con la distribución de *"conversion”*para cada una de las categorías de “*platform*".

	0	1
Accesorios para Vehículos	0.90417981	0.09582019
Agro	1.00000000	0.00000000
Alimentos y Bebidas	0.94554328	0.05445672
Animales y Mascotas	0.96132597	0.03867403
Antigüedades y Colecciones	0.85024155	0.14975845
Arte, Librería y Mercería	0.97069168	0.02930832
Bebés	0.93221194	0.06778806
Belleza y Cuidado Personal	0.90556937	0.09443063
Cámaras y Accesorios	0.92722372	0.07277628
Celulares y Telefonía	1.00000000	0.00000000
Celulares y Teléfonos	0.85167502	0.14832498
Computación	0.93294919	0.06705081
Consolas y Videojuegos	0.85552941	0.14447059
Deportes y Fitness	0.91720512	0.08279488
Electrodomésticos	1.00000000	0.00000000
Electrodomésticos y Aires Ac.	0.89653347	0.10346653
Electrónica, Audio y Video	0.82963353	0.17036647
Eletrodomésticos	1.00000000	0.00000000
Esportes e Fitness	1.00000000	0.00000000
Herramientas y Construcción	0.91154952	0.08845048
Hogar, Muebles y Jardín	0.91454951	0.08545049
Industrias y Oficinas	0.92891882	0.07108118
Instrumentos Musicales	0.92300557	0.07699443
Joyas y Relojes	0.92560175	0.07439825
Juegos y Juguetes	0.92522495	0.07477505
Libros, Revistas y Comics	0.96800000	0.03200000
Música, Películas y Series	0.95000000	0.05000000
Otras categorías	0.98412698	0.01587302
Ropa y Accesorios	0.93395455	0.06604545
Ropa, Bolsas y Calzado	1.00000000	0.00000000
Salud y Equipamiento Médico	0.90953422	0.09046578
Souvenirs, Cotillón y Fiestas	0.85329341	0.14670659



Si hacemos una tabla con la proporción de “*conversion”* para cada categoría de “*platform”,* el resultado es el siguiente:

	0	1
/mobile/android	0.91583005	0.08416995
/mobile/ios	0.90369445	0.09630555
/web/desktop	0.87239624	0.12760376
/web/mobile	0.91730549	0.08269451

Podemos notar que la proporción de artículos para los cuales hubo compras cambia dependiendo de la plataforma.

Variables creadas.

La primer variable que creamos está relacionada con si el producto tiene garantía o no. En el dataset original, la variable “*warranty”* no era binaria; nosotros decidimos hacerla binaria. Para hacerlo, le otorgamos un 0 a las observaciones que no tenían información o decían “Sin garantía”; y le otorgamos un 1 al resto.

Por otra parte, decidimos hacer análisis de sentimiento para el titulo de la publicación, cuya información está en la variable "*title*" del *dataset* original. Para ello, hicimos *Bag of Words* con el texto del título, quedándonos con las palabras que aparezcan al menos en 3000 observaciones. Entre las palabras que aparecían, observamos algunas interesantes como "oferta", "cuotas", "original", etc.; pensamos que quizás podrían tener poder explicativo.

Finalmente, hicimos *One Hot Encoding*; creando una columna para todas las categorías de productos, "*platform*", "*logistic_type*", y "*listing_type_id*" (indica si el item es *gold_pro* o *gold_special*), y para cada una de las palabras que aparecen en al menos 3000 títulos.

Modelo implementado.

Antes de correr el modelo, convertimos nuestro *dataset* de formato *data.frame* a formato *sparse*, dado que el mismo tenía muchas columnas de unos y ceros.

En cuanto al modelo, utilizamos la librería LightGBM, la cual sí acepta datos en formato *sparse*, y creamos un árbol de decisión, buscando los mejores hiperparámetros para el mismo a través de *random search*.

Conjunto de validación.

En primer lugar, separamos el conjunto de datos original en *training* y *test*. El *training set* son las observaciones para las cuales hay NA en "*Row_ID*". El *testing set* son las observaciones restantes.

Del *training set*, extrajimos de forma aleatoria 1800 observaciones, las cuales serán el *validation set*.

Scores obtenidos.

De todos nuestros intentos en la competencia de Kaggle, el máximo *score* que pudimos obtener fue de 0.89407 en el *Public Score* y 0.87800 en el *Private Score*.