UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Minería de Datos

Primer Avance: Proyecto Integrador

Equipo 13

Grupo 002, Miércoles

1812678, López Ávila, Gilberto Noé

1810699, Ortiz de la Rosa, Vanessa

1801990, Ovalle Blanco, Miguel Eduardo

28 de octubre de 2020

1- Título de la base de datos

Sales of summer clothes in E-commerce Wish
URL: https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish

2- Descripción de los datos

La base de datos está compuesta por tres archivos de tablas en formato CSV, de las cuales consideraremos un conjunto de columnas del archivo principal. Las siguientes columnas contienen información útil para nuestros objetivos. Estas columnas están contenidas en el archivo **summer-products-with-rating-and-performance_2020-08.csv**.

title: Título del producto para la configuración francesa de Wish; original en inglés o traducción en francés (cadena).

title orig: Título original en inglés del producto (cadena).

price: Precio de venta del producto (flotante).

retail price: Precio original del producto, diferente al precio de venta si el producto tiene un descuento (flotante).

units sold: Aproximación del número de unidades vendidas (entero).

uses ad boosts: Si el vendedor decidió pagar por rezaltar su producto en la página (booleano).

rating: Calificación promedio del producto entre 1 y 5 (flotante).

rating count: Total de calificaciones del producto (entero).

badge local product: Si cuenta con la insignia de recogida inmediata para que el cliente pueda ir por el producto (booleano).

badge product quality: Si cuenta con la insignia de producto verificado el cual es destacado por recibir contantemente buenas reseñas (booleano).

badge fast shipping: Si cuenta con la insignia de envío rápido que indica que el producto tiene menos tiempo de espera (booleano).

tags: Etiquetas asignadas por vendedor, separadas por comas (cadena).

product color: Color principal del producto (cadena).

product variation inventory: Tamaño del inventario, con cuenta máxima de 50 (entero).

shipping option name: Opción de envío (cadena).

shipping option price: Costo extra de la opción de envío (flotante).

shipping is express: Si la opción de envío es express (booleano).

countries shipped to: Número de países para los cuales está disponible el envío (entero).

has urgency banner: Si el producto tiene una bandera de urgencia (booleano).

urgency text: Texto de la bandera de urgencia para los productos con bandera (cadena).

origin country: País de origen del producto (cadena).

merchant info subtitle: Información extra sobre el vendedor, como el porcentaje de críticas positivas, sin editar (cadena).

merchant rating count: Total de calificaciones del vendedor (entero).

merchant rating: Calificación promedio del vendedor entre 1 y 5 (flotante).

merchant id: Identificador único del vendedor (cadena).

merchant has profile picture: Si el vendedor tiene foto de perfil (booleano).

product id: Identificador único del producto (cadena).

Además, consideramos las columnas del archivo **unique-categories.sorted-by-count.csv**.

count: Número de veces que aparece la palabra de la columna keyword en la columna de tags del archivo principal (entero).

keyword: Palabra clave dentro de la columna tags del archivo principal (cadena).

Omitiremos las siguientes columnas del análisis ya que no presentan información de interés para resolver el problema, su información se encuentra resumida en otra columna, o su implementación requiere métodos más complejos.

currency buyer: Moneda del precio (cadena). Omitida pues todos los precios están en euros (EUR).

rating five count, rating four count, rating three count, rating two count, rating one count: Total de calificaciones de 5, 4, 3, 2 y 1 estrellas, respectivamente (entero). Omitidas al tener un resumen de estas columnas en calificación promedio.

badges count: Total de insignias otorgadas por Wish al producto (entero). Omitida ya que esta columna cuenta las insignias ya incluidas en sus columnas respectivas.

product variation size id: Una de las tallas del producto (cadena). Omitida ya que solo muestra una de las varias tallas disponibles al comprar el producto.

inventory total: Total del inventario para todas las variaciones del producto (entero). Omitida ya que los inventarios son mayores a 50 y la columna solo muestra el valor 50.

merchant title, merchant name: Nombre público y de usuario del vendedor del producto, respectivamente (cadena). Omitidas pues no consideraremos las características de los nombres de otros vendedores.

merchant profile picture. product url, product picture: URL de la foto de perfil del vendedor del producto, del producto, y de la imagen principal del producto, respectivamente (cadena). Omitidas ya que para trabajar con las imágenes se requiere acceder a la dirección y utilizar técnicas de análisis de imágenes.

theme: Palabra usada en el buscador de Wish para encontrar este producto (cadena). Omitida pues todos los productos se encontraron al buscar “summer”.

crawl month: Mes de búsqueda del producto (cadena). Omitida pues todos los productos se consultaron el 1 de agosto de 2020.

También omitimos el archivo **unique-categories.csv**, pues la información de estas categorías se muestra en unique-categories.sorted-by-count.csv.

3- Justificación del uso de los datos

Entre las características que nos convencieron para usar esta base de datos es la cantidad abundante de información que posee en columnas, el formato de CSV que tiene y su cantidad de datos que en total son 1573, de igual forma se tiene por separado otro archivo con las etiquetas y el número de veces con que aparece. Esto a nuestro parecer nos permite una facilidad para controlar los datos y nos apoya a entenderlos de una mejor manera para realizar el trabajo.

Así mismo, en la actualidad, las compras en línea es una tendencia que va creciendo conforme pasa el tiempo como parte de una era más digitalizada por lo que el que nosotros interactuemos con una página como lo es Wish, nos es familiar a otras que se dedican a esto, con lo cual teniendo ya conocimiento de ello pues el saber el significado de datos y darles interpretación es un punto a favor para abordar una problemática como esta ya que en un futuro próximo será algo aún más normal.

4- Planteamiento del problema

Un emprendedor cuenta con una línea de productos de verano que vende a través de varios locales de su ciudad, está interesado en hacer crecer su negocio y que sus productos puedan llegar a más partes del mundo.

Le recomendaron incursionar en las ventas online a través de la plataforma de Wish pero tiene miedo de fracasar en el intento y hacer una sobreproducción de sus productos. El quisiera saber cuáles de sus productos podría comerciar en esta plataforma de tal manera

que resulten productos exitosos, es decir que tengan una buena venta y sean populares entre los usuarios.

Por lo tanto, el problema que notamos tiene que ver con distintos aspectos que se manejan en la plataforma como el uso de publicidad, el tipo de envío que se maneja, el vendedor que se ubica entre el gusto del usuario y el producto, entre otras cosas determinantes como lo es en una temporada fuerte que es el verano. Así que sería de buen provecho el saber lo determinante que, desde el punto de vista del cliente, mejoraría su experiencia en las compras online aumentado el gusto de hacerlo y con ella crear mejores relaciones para futuras compras.

De esta manera, nuestro enfoque es que a partir de los datos que tenemos de las ventas del verano, ayudar al emprendedor que quiera vender en la plataforma de Wish para que esté preparado el siguiente verano y que tenga el resultado que espera en su negocio.

5- Objetivo final

Objetivos principales:

- Obtener una serie de características que hacen que un producto sea exitoso en la plataforma de Wish.
- Predecir las unidades vendidas que tendrá un producto de acuerdo a las características que lo componen.
- Dar una serie de recomendaciones al vendedor para lograr un aumento de sus ventas.

Objetivos secundarios:

- Realizar un análisis descriptivo sobre las características de los productos de verano de Wish
- Clasificar los productos mostrados en categorías de características en común y determinar el éxito de estas categorías.
- Reconocer características similares de los productos más vendidos para identificar el tipo de productos que aparecerá más frecuentemente en las recomendaciones de estos productos exitosos.

6- Planeación de las herramientas a utilizar

Utilizaremos herramientas de predicción para determinar el éxito en ventas de un producto a partir de otras variables. Ordenamos las siguientes técnicas de minería de datos de acuerdo con qué tan útiles esperamos que sean para llegar a estos objetivos.

1. **Regresión Lineal Múltiple:** Al tener una variedad de datos de distintas características numéricas, podremos determinar un subconjunto de columnas significativas, determinar su influencia en el modelo de regresión, y construir una ecuación capaz de predecir la cantidad de unidades vendidas para un producto dadas las características especificadas.
2. **Árboles de decisión:** Es posible utilizar árboles de regresión, o incluso bosques aleatorios, para determinar las unidades vendidas a partir de varios criterios de decisión basados en las características más importantes del producto.
3. **Algoritmos de clasificación:** En vez de predecir una variable numérica, podremos construir una clasificación del éxito de productos de acuerdo con un conjunto de indicadores, y utilizar estos algoritmos para determinar si algún nuevo producto, según sus características, pertenecerá a la categoría de productos exitosos.