Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey

Materia

Inteligencia artificial avanzada para la ciencia de datos II

Tarea

Modulo 2 - 7.- Feature Selection

Estudiantes

Cleber Gerardo Pérez Galicia - A01236390

Juan Pablo Bernal Lafarga - A01742342

Jacobo Hirsch Rodríguez - A00829679

Eryk Elizondo González - A01284899

Profesor

Selección de Features

Para el proyecto de predicción de éxito en productos de nuevo lanzamiento de la empresa Arca Continental, se consideraron las siguientes variables que, creemos, son las más relevantes para definir el "ADN" de un producto, maximizando la capacidad predictiva y/o explicativa de acuerdo con nuestro algoritmo de similitud cosenoidal entre productos. Dicho ADN contiene:

- [Productos_Por_Empaque]
- [MLSize]
- [Returnability]
- [Size]
- [GlobalFlavor]
- [Container]
- [GlobalCategory]
- [BrandGrouper]
- [Presentation]
- [Brand]

Las variables cubren desde el número de contenedores por empaque, el volumen de los productos, si es retornable, el tamaño como "familia" o "individual", el sabor del producto, el material del contenedor, el tipo de producto, si el producto es un líquido o comida, la marca a la que pertenece y el grupo de dicha marca.

Todas estas características en conjunto vuelven a los productos únicos, diferenciables uno del otro dentro de un espacio vectorial.

Métodos y Técnicas Utilizadas

Siguiendo la idea del Análisis de Componentes Principales (PCA), redujimos la base de datos de los productos a la menor cantidad de columnas que nos ayudan a diferenciar todos los productos y que explican la mayor información del mismo. Esta selección manual se hizo debido a que la información de las otras columnas era similar entre sí, siendo una combinación textual de las features seleccionadas con información irrelevante o desconocida por nosotros y en diferentes posiciones.

Vector Embeddings

como segunda técnica de reducción de dimensionalidad utilizamos la técnica de vector embedding, esta técnica se utilizó considerando que la mayoría de las columnas de nuestro

dataset principal productos son variables categóricas, descripciones textuales de los productos que vende Arca continental, esta técnica nos permitiría capturar la similitud semántica entre columnas. Este enfoque permite medir el grado de relación entre los textos de diferentes columnas, incluso cuando no utilizan las mismas palabras exactas. La técnica elegida fue la similitud de coseno aplicada a los embeddings, la cual proporciona un valor entre -1 y 1 que indica qué tan alineados están los significados de dos textos.

Además de esta técnica se utilizo la similitud de Jaccard para medir la coincidencia exacta de vocabulario entre columnas. Esta técnica compara los conjuntos de palabras en cada columna y calcula la proporción de palabras compartidas entre ambas. La similitud de Jaccard fue particularmente útil para detectar columnas con vocabularios idénticos o muy similares en sus términos específicos. Ayudando a distinguir la similitud semántica proporcionada por los embeddings a la similitud léxica proporcionada por la similitud de jaccard. Teniendo estos dos criterios pudimos identificar las columnas que eran redundantes en estos dos aspectos

En el procedimiento, Para cada par de columnas, se calculó la similitud de coseno entre sus vectores de embeddings, obteniendo así un valor de similitud que refleja el grado de relación semántica entre ellas.
Estos valores se organizaron en una matriz de similitud semántica, donde cada celda representaba el nivel de similitud de coseno entre dos columnas específicas. Esta matriz permitió identificar columnas con relaciones semánticas altas.

Paralelamente, se generó una representación Bag of Words para cada columna, que posteriormente se transformó en un vector binario. En este vector, cada posición indicaba la presencia o ausencia de palabras específicas en la columna. A continuación, se calculó la similitud de Jaccard entre los vectores binarios de cada par de columnas, lo que permitió cuantificar la coincidencia exacta de vocabulario entre ellas. Estos resultados se organizaron en una matriz de similitud léxica, donde cada celda contenía el valor de similitud de Jaccard entre dos columnas. Esta matriz facilitó la identificación de columnas con vocabularios muy similares o idénticos.

Criterios Adicionales

Desconocemos exactamente los pesos de cada feature en la toma de decisión de compra del producto por cliente. Aunque sabemos que sí que existen diferencias entre las propias categorías de las variables. Siendo que los productos de la marca “Coca Cola” probablemente venden más que los productos de la marca “Joya”, y dado el análisis del perfil de compras de los clientes, se está generando un sesgo donde cada cliente muestra su preferencia hacia cierta marca, sabor, tamaño del producto, el material, entre otros factores.

Se consideraron criterios adicionales al decidir qué columna conservar cuando se encontraban redundancias. Entre estos criterios se incluyeron:

- **Variedad de valores únicos**: Se priorizaron las columnas con mayor número de valores únicos, ya que representan una mayor capacidad de discriminación.
- **Distribución de valores**: Se evaluó la uniformidad de la distribución de valores en cada columna, conservando aquellas con una distribución más equilibrada.

Resultados

Para los pares con una similitud de coseno superior a un umbral de 0.95 se consideraron semánticamente redundantes y para la similitud de jaccard se probaron distintos umbrales de similitud de Jaccard, comenzando desde 0.8 y reduciéndolo gradualmente a 0.5, Un umbral más bajo permitió identificar redundancias adicionales al exigir menos coincidencia exacta en vocabulario.
terminamos seleccionando las siguiente columnas para ser eliminadas con los criterios seleccionados:
BrandGrouper, Pack, ProductType, SegAg, SegDet