Priorización de Clientela con Aprendizaje Automático

Inteligencia Artificial Avanzada para la Ciencia de Datos II (Gpo 101)

Pixel Duck
Instituto Tecnológico y de Estudios Superiores de Monterrey
Cleber Gerardo Pérez Galícia, Juan Pablo Bernal Lafarga, Jacobo Hirsch Rodríguez, Eryk Elizondo González
Noviembre 2024

Introducción

En el competitivo entorno del mercado, el análisis predictivo se ha convertido en una herramienta crucial para el éxito de las empresas, ya que permite analizar grandes volúmenes de datos históricos y encontrar patrones que permitan realizar predicciones de ventas alrededor de nuevos productos.

Problemática

Arca Continental, la segunda embotelladora de Coca-Cola más grande de América Latina, busca desarrollar un algoritmo de inteligencia artificial capaz de predecir el éxito de nuevos productos en el mercado, basándose en el comportamiento de venta de sus clientes y sus características demográficas.

Herramientas y recursos

Utilizamos Python en un ambiente local en Visual Studio Code para mantener privados los datasets CSV proporcionados por Arca Continental:

- 'ventas.csv': Alrededor de 2 millones de registros de ventas del 2019-2022. Con el producto, cliente y cantidad de la compra.
- 'products.csv': Aproximadamente 800 productos y columnas con información textual de sus características.

También hicimos uso de varias librerias de python, como Pandas para el manejo y la trasnformación de datos, Numpy para operaciones de arrays multidimensionales y Scikit-Learn para operaciones de Machine Learnig e IA.

Proceso de Ciencia de Datos

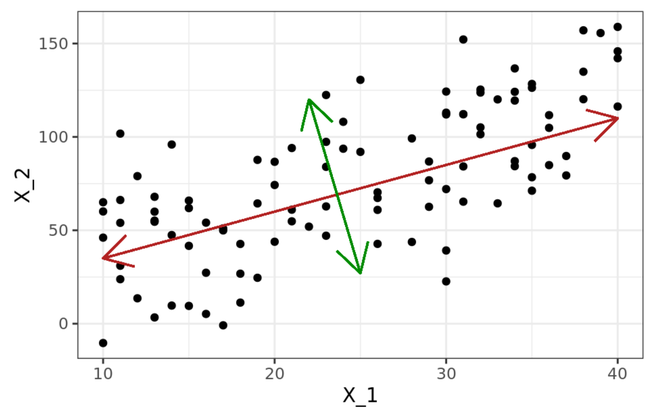
Limpieza de Datos

Para la limpieza, se identificó redundancia entre las variables categóricas que identifican a los productos de la base de datos 'products', y con el objetivo de reducir la dimensionalidad de las tablas, se utilizaron 2 técnicas de procesamiento de lenguaje natural con el fin de identificar las similitudes semánticas entre las categorías. Las 2 técnicas son:

- Vector Embeddings
- Similitud de Jaccard

Transformación de Datos

Para la transformación de los productos, se aplicó la función de One-Hot Encoding de Scikit Learn en las columnas categóricas para convertirlas a binarias. Luego, se normalizaron los valores de las columnas utilizando la función de Min-Max Scaling. Finalmente, se redujo la dimensionalidad con la técnica de Análisis de Componentes Principales.



Para la transformación del dataset 'ventas', se reformateó la variable de la fecha, la cual se encontraba como tipo entero, a un datetime. Luego, se calculó la frecuencia de ventas de los productos por cliente al obtener una proporción entre los meses donde hubieron ventas entre el total de meses a evaluar de todos los productos. Finalmente, se obtuvo la magnitud de ventas relativas de los productos al obtener la proporción de todos los galones vendidos por producto por cliente entre la venta de galones totales de todos los productos del mismo cliente.

Modelo de Solución

Definición de Nuevo Producto

Para la solución de la problemática primero se debe definir un nuevo producto a comparar el cual posee un listado de características que describen al producto. Este producto es la entrada para el modelo, el cual usará los datos previos como entrenamiento para poder priorizar la clientela con base a su compatibilidad con el nuevo producto.

Una vez que este producto se define, se transforman las columnas del producto de forma igual a los datos previos, permitiendo que se puedan comparar entre si.

Similitud Coseno

Una vez que se obtiene el nuevo producto, se obtiene la similitud coseno entre el nuevo producto y todos los productos en la base de datos, con el objetivo de definir numéricamente la similitud entre los productos.

$$\text{SimCos}(u, v) = \frac{u \cdot v}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Donde:

- $\mathbf{u}$ es el producto nuevo
- $\mathbf{v}$ cualquier otro producto

Compatibilidad

Una vez que se obtiene la similitud coseno, se une con la frecuencia y magnitud de ventas calculadas previamente de forma que se obtienen 3 indicadores, estos se les multiplica un peso personalizado para poder determinar la influencia de cada indicador en la determinación de la compatibilidad del producto.

$$compatibility = \alpha \times SF + \beta \times SP + \gamma \times PS$$

Donde:

- $\alpha =$ Peso de la frecuencia de ventas,
- $\beta =$Peso de la proporción de ventas,
- $\gamma =$ Peso de la similitud de productos.

Evaluación y Resultados

Al eliminar temporalmente un producto del historial de ventas, podemos predecir el listado de clientes compatibles y comparar dicha lista con los datos reales, de tal manera que podemos crear nuestras propias métricas de precisión con base en los verdaderos positivos de tal manera que obtenemos:

```
Precisión Top 1: 0/1 = 0.00%
Precisión Top 5: 1/5 = 20.00%
Precisión Top 10: 1/10 = 10.00%
Precisión Top 20: 7/20 = 35.00%
Precisión Top 50: 22/50 = 44.00%
Precisión Top 100: 42/100 = 42.00%
```

Donde el 'Top n'denota los n mejores clientes reales, y el porcentaje de estos clientes que fue predicho correctamente por el modelo. La precisión promedio obtenida al probar 7 productos con los parámetros $\alpha = 0{,}2, \beta = 0{,}2, \gamma = 0{,}6$ en el modelo fue de aproximadamente 40 %. Es decir, 40 de 100 clientes predichos por el modelo compran los nuevos productos por su parecido con otros.

Conclusiones

El modelo propuesto combina métricas clave y técnicas avanzadas para predecir el éxito de nuevos productos, facilitando decisiones estratégicas al identificar clientes con mayor probabilidad de compra, optimizando recursos y estrategias de marketing. Además, la evaluación validó su precisión al predecir compradores prioritarios, demostrando su utilidad práctica. Se podría mejorar con un tipo de retroalimentación basado en penalizaciones y recompensas para modificar las constantes de Compatibilidad que afectan la precisión del modelo, por lo cual es importante actualizar constantemente el modelo para incrementar aún más su eficacia.