Resumen del TFM: Clusterización de Establecimientos y Aplicaciones al Negocio

Marta Sola Corporales
Máster en Técnicas Estadísticas e Investigación Operativa
Universidad de Santiago de Compostela

Enero 2026

Resumen

Se hace constar que el Trabajo de Fin de Máster titulado *"Clusterización de Establecimientos y Aplicaciones al Negocio"* no cuenta con la autorización para su publicación en abierto, ni para su difusión a través de los repositorios institucionales de la universidad.

Esta restricción obedece a los acuerdos de confidencialidad suscritos con la entidad colaboradora, Hijos de Rivera S.A.U.. El presente trabajo utiliza datos reales de ventas, censo de clientes y variables estratégicas del canal HORECA cuya divulgación podría comprometer la ventaja competitiva de la empresa y vulnerar la normativa de protección de datos. A continuación, se expone un resumen detallado de la metodología y resultados sin revelar información sensible.

1. Introducción y Definición del Problema

El principal desafío abordado en este trabajo consiste en la segmentación del canal HORECA con el objetivo de establecer una jerarquía clara entre clientes y definir segmentos diferenciados sobre los que aplicar estrategias de negocio específicas.

La heterogeneidad del sector hace necesario agrupar los negocios en función de sus patrones de venta y características estructurales. Esta agrupación permite identificar perfiles de comportamiento similares, lo que facilita la unificación de estrategias comerciales. De este modo, se busca optimizar la asignación de recursos en campañas de marketing y promociones, evitando acciones genéricas y enfocando los esfuerzos en las necesidades y potenciales reales de cada grupo de establecimientos.

Además, se exploran diferentes aplicaciones al negocio. Por un lado, se ejecuta un modelo de movilidad entre clusters para locales hacia otros de mayor valor. Por otro lado, se plantea un modelo que solventa la dificultad de identificar clientes potenciales en el mercado que se asemejen a los clientes actuales, con el fin de optimizar las visitas comerciales y aumentar la tasa de conversión.

2. Fuentes de Datos y Preprocesamiento

Para la realización del estudio hemos utilizado tres fuentes de datos principales:

- **Datos Internos**: Histórico de ventas desagregado por referencia, familia de productos y envase.
- **Datos de Infraestructura**: Instalaciones técnicas de los locales, distribuidores, delegaciones comerciales etc.
- **Censo de Mercado Externo**: Información de la totalidad de locales de hostelería de España, incluyendo los locales no clientes, que consta de variables geográficas así como relativas a la infraestructura y al funcionamiento del local. Dicho censo está vinculado mediante un identificador con los locales propios de Hijos de Rivera.

El preprocesamiento de los datos supuso uno de los mayores retos del trabajo debido a la alta dimensionalidad, la presencia de datos nulos y la existencia de valores atípicos, típicos en distribuciones de ventas asimétricas. Adicionalmente, se requirió de un complejo proceso de saneamiento y vinculación de locales clientes con el censo, esencial para corregir enlaces erróneos y asociar correctamente locales del censo externo que carecían de identificadores comunes en los sistemas internos.

3. Análisis No Supervisado: Clusterización

El núcleo del trabajo se centra en el descubrimiento de patrones latentes mediante técnicas de aprendizaje no supervisado. Tras evaluar diferentes algoritmos de clusterización así como diferentes distancias, k-means y métodos jerárquicos se descartaron debido a su sensibilidad a datos atípicos, a la dimensionalidad del dataset o al coste computacional que requerían.

La metodología seleccionada fue el algoritmo CLARA (Clustering Large Applications). Este método, basado en la búsqueda de medoides (elementos representativos reales de la muestra), demostró una robustez superior frente a los outliers y a la alta dimensionalidad dada por el amplio catálogo de productos. Esto último se debe a que el algoritmo CLARA permite seleccionar distancias diferentes a la euclídea, que sufre de la maldición de la

dimensionalidad. La distancia Manhattan captura mejor las diferencias estructurales en altas dimensiones.

Partiendo de un rango de clusters establecido por la empresa, la validación del número óptimo de clusters se realizó mediante el criterio de la Silueta Media, estadístico Gap y la interpretación de negocio. El resultando en una segmentación final de 17 clusters. Esta estructura reveló una jerarquía clara definida por el tipo de instalación de los locales y el portfolio de productos.

4. Movilidad entre Clusters y Desarrollo de Clientes

Una vez definida la segmentación de locales, se desarrolló un modelo de movilidad entre clusters. El objetivo no es solo clasificar al cliente, sino trazar una hoja de ruta para su desarrollo hacia clusters de mayor categoría. Esta estrategia comercial se focalizó en Madrid en el barrio de Malasaña.

Para ello, se construyó un sistema de recomendación basado en grafos de transición que calcula la probabilidad condicional de que un cliente evolucione de un cluster de menor valor a uno de categoría superior, en función de los productos recomendados. Este análisis permitió identificar artículos cuya introducción en el punto de venta cataliza el crecimiento y fidelización del cliente. De esta forma, el modelo se traduce en una estrategia de venta personalizada para maximizar el valor del cliente.

Para estimar estas probabilidades, se implementó el algoritmo XGBoost. La elección de este modelo nace de la necesidad de integrar y validar la **hipótesis del Contagio Comercial**, la cual postula que el comportamiento de compra de un local no es un evento aislado, sino que está influenciado por la oferta de su entorno inmediato (lo que venden los vecinos afecta a la demanda propia). Aunque existe un modelo de recomendación propio de la empresa, este trabajo pretendía explorar esta hipótesis e integrarla en el barrio de Malasaña.

El análisis de importancia de variables del modelo determinó que efectivamente, las variables de entorno que miden la presencia del producto en la zona son determinantes, sufriendo los locales un claro efecto de contagio comercial.

5. Captación de Locales y Solución al .ªrranque en Frío"

Como extensión natural del sistema anterior, la metodología se adaptó para resolver el problema del ”Arranque en Frío": el reto de inferir el potencial comercial de un local del que no se dispone de histórico de ventas.

En una primera fase, se desarrolló un modelo Random Forest con balanceo de clases para distinguir entre clientes potenciales y no potenciales. El análisis de importancia de variables de este modelo preliminar arrojó un hallazgo clave: la variable con mayor poder predictivo fue el porcentaje de locales vecinos que ya son clientes. Esto confirmó nuestra hipótesis de contagio comercial: cuando tus vecinos venden cierta marca, existe una presión de demanda que hará más probable que la acabes comprando.

A continuación, se procedió a adaptar el modelo de recomendación utilizado en la fase de movilidad. Para ello, se eliminaron las variables relativas al histórico de ventas y se reentrenó el algoritmo utilizando exclusivamente datos de infraestructura y la presión comercial de cada marca en la zona. Si bien esta adaptación implicó sacrificar parte de la sensibilidad original, permitió solucionar eficazmente el problema del arranque en frío.

Por un lado, tenemos una hoja de ruta priorizada sobre qué locales son más propensos a ser captados y, por otro, un sistema de recomendación específico para los mismos. Para garantizar su uso operativo, estos resultados se introdujeron en una capa de visualización en Google Maps, permitiendo al equipo comercial disponer en sus móviles de una guía de captación inteligente y fácil de usar sin conocimientos técnicos.

6. Conclusiones

El trabajo concluye que la aplicación de técnicas estadísticas avanzadas permite facilitar estrategias comerciales y optimizar esfuerzos. La segmentación propuesta ofrece un lenguaje común para clasificar a los clientes más allá de su facturación. Por otro lado, las estrategias de negocio propuestas dotan al equipo comercial de una herramienta que facilita la captación de nuevos locales y la movilidad de los clientes a categorías superiores.