



Cajamar UniversityHack 2022

Atmira Pharma Visualization

Universidad Complutense de Madrid

Datahackers

Daniel Piriz, Alejandro Vidal, Javier Pardo



INFORME FINAL CREADO

[Ver el dashboard de visualización](#)

PROCESO Y METODOLOGÍA

Con el fin último de generar información útil de cara a mejorar la toma de decisiones por parte de Atida, el proyecto se ha enfocado en crear una visualización avanzada que permite comprender mejor los datos detrás de la plataforma y la generación de modelos predictivos que robustecen la calidad de los insights generados.

Lo primero que se hizo fue un análisis exploratorio de los datos. Al encontrar valores perdidos y mal codificados se realizó una inicial depuración y limpieza de los datos. Pensando en optimizar la visualización en mapas y con mayor detalle se realizó un webscrapping para relacionar los zipcodes y ciudades con su correspondiente región y país. Con esto, se comenzaron a crear los dashboards en PowerBI y de forma paralela se generaron modelos predictivos de reincidencia de compra y de recomendación de productos que posteriormente se incluyeron en la visualización. Finalmente, en la etapa final, se creó la página web para facilitar la visualización del trabajo realizado.

Es destacable el uso de *Git* para controlar el avance en las versiones de software y la contribución de cada uno de los miembros del equipo en las diferentes etapas del proceso. Dentro de la utilización de esta tecnología, cada componente del equipo realizó su desarrollo en ramas distintas que, posteriormente, se unificaron en la rama *main* o principal. Para la comunicación diaria se empleó la herramienta *Discord* con el fin de transmitir información de una manera más directa a través de imágenes, chat o llamada de voz. Además, se programaron reuniones semanales telemáticas para gestionar los avances y problemas detectados. También se realizaron reuniones presenciales para abordar las directrices generales y objetivos prioritarios del proyecto.

En lo referente a la metodología, cada miembro lideró una tarea concreta:

- Daniel Piriz lideró la vertiente más estadística, con la depuración de datos, la construcción del modelo *XGBoost* de fidelización de clientes y la página web de Wordpress mediante la herramienta de *Elementor*, participando de forma activa en la elaboración del recomendador.
- Javier Pardo lideró el apartado de visualización gracias a sus conocimientos previos de *PowerBI*. Añadido a ello, fue un soporte vital y fundamental para el *storytelling* y la capacidad visual del proyecto.
- Alejandro Vidal lideró la vertiente más técnica. Así, debido a su experiencia, fueron posibles los *webscrapping* geográficos y de imágenes efectuados en “Google Cloud Platform” (GCP). Por otro lado, también tuvo una alta implicación en la elaboración del recomendador y en la coherencia técnica general del proyecto.

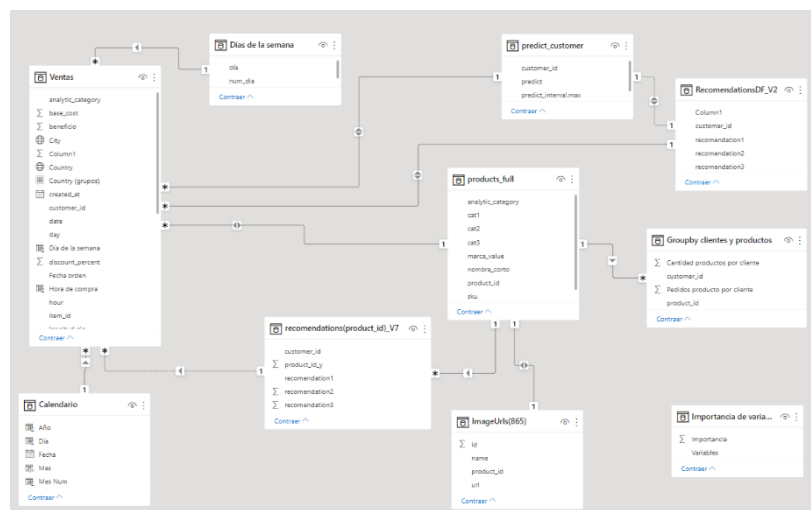


TÉCNICAS APLICADAS Y RESULTADOS OBTENIDOS

El análisis exploratorio de los datos se realizó con lenguaje *Python* y su librería *Pandas* usando “Jupyter Notebooks”. Este análisis previo mostró unas tablas en estado mejorable y que se debían de tratar correctamente antes de su utilización en visualización y modelización. Entre las acciones de limpieza se puede destacar:

- Selección de variables utilizables.
- Generación de nuevas variables.
- Gestión de duplicados.
- Limpieza de zipcodes.
- Obtención de información añadida mediante *webscrapping*.
- Gestión de valores perdidos (imputación por *randomforest*).
- Generación de un dataframe final unificado.

Para generar los modelos se usaron dataframes específicos que incluyen las variables necesarias de acuerdo con su objetivo. Los dataframes de productos y categorías se unieron en uno solo, que junto con el de ventas fueron la base de la visualización. Cada *notebook* utilizado durante la depuración recibe el *dataframe* resultante del anterior. Se han adjuntado, por tanto, el csv original de las ventas de dos años y el requerido por el modelo.

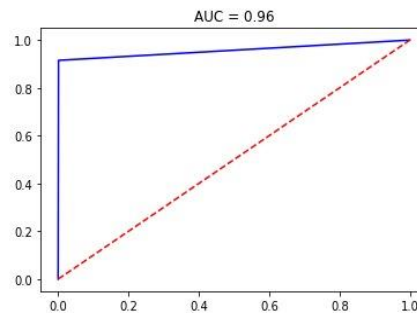


Adicionalmente, se ha empleado la técnica del *webscrapping* con *Selenium* para obtener más información de manera externa a la proporcionada originalmente. De este modo, a partir de la columna *zipcode* se ha limpiado la columna de ciudad y se ha aportado el país y la región adicionalmente. Esta información es relevante de cara a la elaboración de nuevas formas de visualización alternativas y ha sido almacenada en una base de datos basada en *SQLite3*. Todo ello ha sido realizado en una instancia de alta capacidad de procesamiento en la plataforma nube de “Google Cloud” y con el apoyo de procesamiento del lenguaje natural (NLP) para comprobaciones más eficaces a través de la técnica de la “tokenización”. De este modo sólo se han reportado 75.869 errores de un total de 960.930 códigos postales analizados.

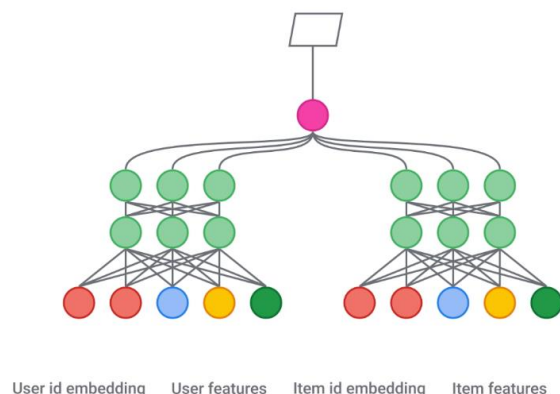


Dado que no se obtuvo valor de los enlaces con imágenes del dataframe recibido, se optó por la técnica del *webscrapping* con *Selenium* para obtener imágenes válidas en la gran mayoría de productos (sólo 889 valores nulos de cerca de 19.000 productos únicos depurados previamente) con una metodología similar a la descrita en el párrafo anterior.

Con éxito, se realizó un modelo en base a clientes, obteniendo qué factores hacen que un cliente vuelva a comprar y qué clientes tienen mayor probabilidad de volver a comprar en miFarma. El algoritmo empleado es *XGBoost* a través de la librería de *SKlearn*. Así, este modelo ha resultado bastante fiable, con un *accuracy* del **0.96** y un AUC del **0.96** también. Además, se mantiene un aceptable grado de sensibilidad y especificidad que confirman las virtudes de dicho modelo.



Por otro lado, se aprovecharon las bondades de *Tensorflow Recommenders* para construir un prototipo de sistema de recomendaciones de compras para cada cliente mediante redes neuronales de perceptrones multicapa que permitiese incluir en un futuro mayor profundidad de features generando un modelo más robusto. Esta red busca similitudes entre productos con el fin de ofrecer la mejor alternativa a cada cliente en función de sus compras realizadas. Actualmente se encuentra en un estado funcional, pero con amplio margen de mejora y, por esta razón, es uno de los principales puntos de mejora que tendrá que afrontar el proyecto a futuro (comentado en apartados posteriores).



En lo referente a la visualización principal, se empleó la tecnología de *PowerBI Desktop* para la creación de dashboards enfocados hacia la utilidad y la fácil comprensión. De esta manera, el proyecto dispone de tres dashboards principales (ventas, productos y clientes) compuestos por diferentes apartados que desglosan toda la información obtenida y sintetizada en las etapas comentadas anteriormente. Adicionalmente, para la publicación en la web se obtuvo una licencia de PowerBI Pro.



Para la visualización en PowerBI se crearon medidas adicionales que permiten mejorar la interactividad entre gráficos y resumir información de forma adecuada, se aprovechó las opciones que ofrece este software con PowerQuery y el lenguaje DAX. Se han utilizado mapas para dibujar las zonas donde están ocurriendo los eventos de interés. También se utilizan gráficas que resaltan otros criterios como temporales y categóricos además de vectores de actuación adicionales como son clientes y productos. Por otro lado, adentrándose en los interiores de los dashboards, se localiza una intrincada red de relaciones entre tablas que permiten añadir interactividad y, sobre todo, profundidad en la información mostrada.

Finalmente, se ha creado una web para facilitar el acceso a los datos y los apartados del proyecto de una forma eficaz y coherente. Para agilizar la creación del back-end se ha empleado Wordpress, mientras que para el front-end se ha empleado Elementor, un constructor visual para Wordpress.

FUENTES EXTERNAS EMPLEADAS

Como se mencionó en el apartado anterior, se añadió información geográfica extra (país, región, localidad) a partir de la variable de *zipcode*. Para ello, a través de lenguaje *Python*, se empleó la librería de *Selenium* para poder acceder al *website*: <https://worldpostalcode.com/> y recoger la información ofrecida tras introducir el *zipcode* del dataset original para el proyecto.

Por otro lado, para mejorar la calidad de la visualización de algunos apartados del *dashboard* principal, se volvió a utilizar *Selenium* para recopilar imágenes de todos los productos disponibles a través de “Google Imágenes”. Para ello, fue imprescindible el uso de “Google Cloud” para agilizar esta tarea tan costosa a nivel computacional.

ACCIONES DE FUTURO

Más allá de una mejor optimización y depuración del trabajo ya realizado, las líneas maestras para el futuro son:

- **Modelo de predicción temporal**

Con los datos actuales no se ha conseguido hacer un modelo de predicción temporal que cumpliera unos estándares de calidad mínimos con modelos ARIMA. Así, se propone realizar un modelo más avanzado con modelos predictivos supervisados como *XGboost*, con el fin de intentar aportar más valor para escenarios como el presente.

- **Clusterización**

Trabajar sobre un modelo de *clustering* para conocer las relaciones presentes entre ciudades/regiones del territorio español, con posibilidad para ampliar el tamaño de la muestra hacia otros países (habría que enriquecer con más *features* para las distintas regiones más allá de territorio español). De esta forma, se pueden establecer grupos que compartan características a las que aplicar estrategias similares, o establecer proyecciones de crecimiento. Muy similar sería aplicar técnicas similares en productos, viendo que tipo de subgrupos se establecen, como, por ejemplo, un posible grupo de productos estrella.

- **Mejora de la red neuronal para las recomendaciones de producto**

Se plantea un desarrollo más avanzado en la implementación del sistema de recomendación, haciendo uso del potencial que ofrece la red neuronal añadiendo nuevas *features* que mejoren su rendimiento y, por tanto, la calidad de las recomendaciones sugeridas.