



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 marcado hoja de ruta la generación de un modelo que aportase valor añadido y una visualización que permita comprender mejor los datos detrás de la plataforma.

Durante el proceso de elaboración del presente trabajo, se trabajó con una metodología muy enfocada hacia herramientas empleadas en el ámbito profesional.

De este modo, en primer lugar, se utilizó “Git” para controlar las versiones de software que se fueran creando con el paso del tiempo. 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.

En segundo lugar, 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. Se utilizó esta modalidad para obtener una aproximación a un servicio de comunicación parecido a la herramienta profesional “Slack” o “Microsoft Teams”

En lo referente a la metodología, cada miembro se encargó de una tarea concreta:

- Daniel Piriz se encargó de la vertiente más estadística, con la fusión final de dataframes, imputación de valores nulos y generación del modelo final.
- Javier Pardo se encargó del apartado de visualización gracias a sus conocimientos previos de Power BI. Añadido a ello, fue un soporte vital y fundamental para el *storytelling*.
- Alejandro Vidal se encargó de la vertiente más técnica. Así, debido a su experiencia, fue posible el *webscrapping* efectuado en “Google Cloud Platform” (GCP). Por otro lado, organizó las tecnologías empleadas en el proyecto y dio soporte en su ejecución.

Además, se programaron reuniones semanales telemáticas para gestionar los avances y problemas detectados.



TÉCNICAS APLICADAS Y RESULTADOS OBTENIDOS

El primer paso a seguir fue el análisis exploratorio de las tablas a tratar con lenguaje Python y Pandas. 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

Además, se intentó obtener valor tratando otras variables como los enlaces a fotos o categorías más específicas, sin mucho éxito.

El data frame final obtenido es la unión de las tres tablas facilitadas para poder trabajar en el presente proyecto, que recoge la información de las ventas y de los productos vendidos. Cada notebook utilizado 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.

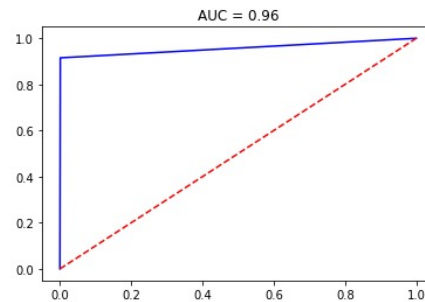
Se ha empleado la técnica del *webscrapping* con Selenium para obtener más información de manera externa a la proporcionada originalmente. De este modo, a partir de la columna *zipcode* se ha limpiado la columna de ciudad y se ha aportado el país y la región adicionalmente.. Esta información es relevante 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.

Por último, es reseñable el empleo de procesamiento del lenguaje natural (NLP) para comprobaciones de palabras más eficaces a través de la técnica de la tokenización.

Los resultados obtenidos con la técnica del *webscrapping* y la *tokenización* son bastante prometedores. De este modo sólo se han reportado 75.869 errores de un total de 960.930 códigos postales analizados como puede apreciarse en la siguiente imagen:



Con el dataframe definitivo, se plantearon dos líneas de modelización con la intención de obtener valor añadido. El primer modelo era una evolución temporal que permitiese hacer estimaciones del volumen de beneficios para futuras semanas, pero con los datos no se consiguió hacer unas predicciones que se pudieran considerar fiables. Por otra parte, con mucho mayor 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. Este modelo es un modelo muy fiable, con una accuracy del **0.96** y un AUC del **0.96** también.



Para que la empresa tenga presente qué está ocurriendo con sus ventas y conocer a sus clientes, se ha realizado un dashboard de visualización interactivo en web mediante Power BI. Se han utilizado mapas para dibujar las zonas donde están ocurriendo los pedidos, dando mayor detalle a España (a nivel de códigos postales), pues es el país con mayor número de ventas. También se utilizan gráficas que resaltan los meses con mayor cantidad de ventas y las horas y días donde más pedidos se realizan a la farmacia.

Al cargar los archivos utilizados en la visualización, se realizó un modelado de datos para relacionar las tablas entre sí. Se añadieron columnas y medidas calculadas a partir de los datos base (algunos como precio total, conteo de los clientes y pedidos únicos y para las ventas totales). Igualmente se agregó una tabla para vincular el número de la semana (del 1 al 7) con el día (de domingo a sábado). Se ajustaron los tipos de los datos según corresponda (por ubicación, número y moneda).

Los filtros que se crean en cada página permiten dar mayor detalle a quien consulte el dashboard. El dashboard busca destacar los productos más pedidos, las marcas y categorías más relevantes, los clientes y los pedidos más grandes para la compañía, con la posibilidad de filtrar por países, considerando que las dinámicas de los clientes y ventas en cada uno pueden ser distintas.

Se ha introducido la información obtenida de la predicción en la aplicación web, de forma que la empresa tenga acceso a la información de forma visual. De entre la información obtenida, cabe destacar que el modelo tiene una gran precisión determinando qué clientes tienen probabilidades de repetir, y que se podrá replicar para nuevos clientes con el objetivo de determinar si son potenciales clientes a ser incluidos en campañas de marketing relacional.

Con esto, la compañía puede mejorar su estrategia de marketing y logística, enfocada en el comportamiento de sus clientes en cada región.



FUENTES EXTERNAS EMPLEADAS


Como se mencionó en el apartado anterior, se añadió información geográfica extra 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.

Posteriormente, esta información recopilada por Selenium es procesada y verificada con el fin de sólo obtener la información relevante.. Dicho proceso de verificación se basa en procesamiento de lenguaje natural (NLP) con la librería “NLTK” y, más concretamente, en técnicas como la tokenización

A partir de este momento, la información extra obtenida es almacenada en una base de datos SQLite3 para ahorrar peticiones innecesarias del mismo zipcode a la web antes comentada(puesto que ya está almacenada en la base de datos).

Todo este proceso es repetido para los 930.960 códigos postales del cuál se compone el dataset de “items_ordered” proporcionado para el proyecto. Sin embargo, este proceso tan costoso a nivel computacional, pudo ser trasladado a una instancia de alta capacidad de procesamiento en Google Cloud para agilizar tiempos.

Configuración de la máquina

Tipo de máquina	c2-standard-4
Plataforma de CPU	Intel Cascade Lake
CPU virtuales para proporción de núcleos	—
	
Dispositivo de visualización	Inhabilitado Habilita esta opción
GPU	Ninguna

ACCIONES DE FUTURO

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

- Visualización de datos más polivalente.
- Generación de distintos modelos que aporten más información.

Para el futuro se plantea un uso más intensivo del Cloud junto con PySpark para poder obtener modelos de mayor complejidad computacional. Por otro lado, se plantea la integración de Power BI sobre un desarrollo web más profesional que permita una interactividad pública más potente.