Informe sprint 2

Proyectos IV: Equipo BI

Proyecto: ERP

Juan Carlos Ávila* Javier Coque† Alejandro García‡ Chantal López§

17, marzo 2024

1 Resumen Sprint 2

En líneas generales, el equipo de BA del proyecto BigEye ha hecho un buen Sprint. No sólo se han hecho las tareas que quedaron pendientes del Sprint anterior—como pueden ser la creación de un dataset para el modelo de segmentación de instancias a partir de datos de Catastro y la creación de un modelo básico de segmentación— sino que se han completado con éxito una amplia mayoría de las tareas propuestas para Sprint.

A lo largo del Sprint, surgieron problemas, que o bien fueron resueltos por ellos mismos, o consiguieron ayuda por parte del equipo de BI —en relación al datalake— o se resolvieron por parte del cliente (Pablo) en la propia auditoría o se quedaron pendientes de resolver en un futuro próximo con este mismo cliente (compatibilidad de Prefect con distintos SO)

*juan.avila@live.u-tad.com

†javier.coque@live-u.tad.com

‡alejandro.gallego@live.u-tad.com

§chantal.lopez@live.u-tad.com

Nota

Juan Carlos, miembro de nuestro equipo, propuso emplear el datalake de Azure como posible solución a su problema. Esto es porque Microsoft aporta más facilidades, en general, al uso de sus herramientas para los estudiantes

2 Puntos Concretos

Habiendo dado un poco de contexto general sobre cómo fue la auditoría, se procede a explicar algunos puntos con algo más en detalle.

Debido a la alta computación por parte del entrenamiento de los modelos, se optó por ejecutarlos en Google Colab. Sin embargo, cada dos horas, esta plataforma detenía su ejecución. Finalmente, se decide hacerlo con la herramienta que ya se estaba usando para MLOps: Prefect.

Esto nos lleva al siguiente punto: problemas con Prefect. Esta herramienta presentaba un problema no resuelto para la auditoría: no se podían gestionar workers desde distintos SO, pues cada uno tenía una ruta distinta y no se pudo establecer *una ruta común* en los ordenadores de cada integrante del equipo auditado. Pablo, el cliente, les comentó que él tiene esa misma situación y que sabe la solución al problema. Entonces, aunque este problema no quede resuelto, se le ha encontrado solución.

Respecto a la obtención de datos para las nubes de puntos (modelo para clasificación una vez empleado el modelo de segmentación), hubo problemas a la hora de obtenerlos. El único recurso proveedor de datos con el que se obtenían buenos resultados, era de pago. El equipo de BA de BigEye optó entonces por descargarse datos ya existentes de una dataset. Con estos datos se obtenía una 51% de *accuracy*; algo que está bastante bien teniendo en cuenta que hay 5 posibles clases en el modelo de segmentación[1].

En cuanto al modelo de TCN, al ser imágenes satélites, la serie temporal presentaba muchos picos (outliers). Esto es debido a los días nublados, en los que se las imágenes captadas son de las propias nubes y no de suelo como tal. Esto provoca que imágenes muy distintas se capten según si el día está nublado o no. Esto, no solo da lugar a una serie temporal con datos que no son *verdaderos*, sino que provoca también que la serie temporal sea mucho más difícil de modelar debido datos tan dispares y

[1] Si es edificio: Residual, Público, Comercial o Religioso. La quinta clase es que no sea un edificio y se clasifique como tal.

aleatorios. Una solución para este problema—que cabe remarcar que nos ha gustado mucho— es la de obtener datos de un sitio en el que, que el día esté nublado sea infrecuente. Este sitio en concreto es el Parque Natural Sierra Alhamilla en Almería. De esta manera, se consiguió un modelo con una mejor precisión.

Algo pendiente para el siguiente Sprint (aunque el equipo auditado ya tiene alguna idea) respecto al punto anterior, es la de determinar qué se quiere hacer con este modelo; qué hacer con los datos obtenidos.

3 Objetivos para el siguiente Sprint

El equipo de BA del proyecto BigEye, hacia el final de la auditoría comentaron los objetivos finales que tienen para el **tercer y último Sprint**. Estos son:

- Como se comentaba recientemente, terminar de dar un enfoque al modelo de TCN.
- Mejorar la precisión (refinar) tanto el modelo de TCN, pero más urgentemente el de segmentación.
- Lanzar a producción estos modelos una vez refinados.
- Consultar con el equipo de BI de BigEye acerca de posibles dudas o de cómo han implementado ciertas cosas que al equipo auditado aún le quedan por hacer o mejorar.
- Crear una interfaz gráfica para mostrar los resultados

4 Comentarios y Consejos

En líneas generales, poco tenemos que decirle al equipo auditado. Están haciendo muy buen trabajo, se nota que han mejorado la comunicación entre ellos—algo que se destacó como mejora en el primer Sprint—, y que lo tienen bastante avanzado. Como se nos mostró en el PowerPoint, tienen en todos los modelos (segmentación, clasificación y predicción) más de un 50% de progreso, llegando incluso al 70% en el último.

5 Conclusión

Aunque tengan que meter un empujón final en este Sprint, han hecho un gran trabajo hasta el momento, y no tenemos ninguna duda de que llegarán al final de Sprint 3 con un producto final adecuado.