



Inteligencia artificial avanzada para la ciencia de datos

Reto

Etapas 5. Evaluación de modelo iteración 2

Diego Arturo Padilla Domínguez - A01552594

Keyuan Zhao - A01366831

Carolina Herrera Martínez - A01411547

Cutberto Arizabalo Nava - A01411431

Jose Pablo Cobos Austria - A01274631

Campus Querétaro

18 de noviembre de 2022

Evaluación de resultados

Importancia de cada variable determinada por el modelo XGBoost

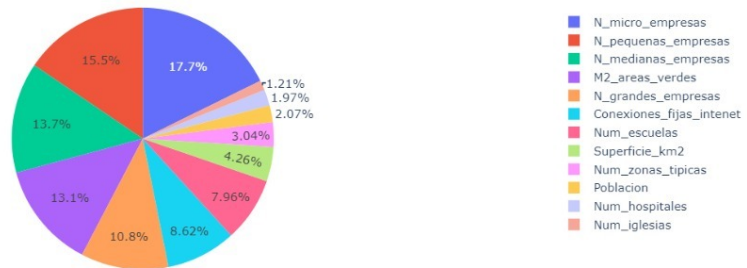


Fig 1. Importancia de cada variable en porcentaje

Los hallazgos que obtuvimos son los siguientes:

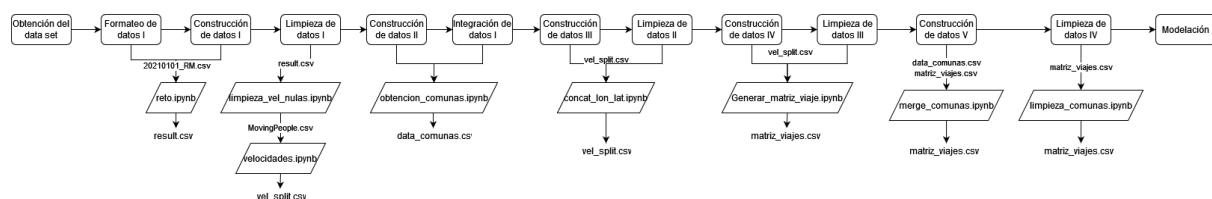
- Hallazgo 1: Observando la Fig.1 podemos observar que **el principal atractor de viajes son las microempresas**, considerando que los pequeños negocios suelen tardar más en adoptar el proceso de transformación digital, es más probable que los empleados de dichos negocios acudan a trabajos presenciales en época de pandemia. Asimismo, vemos que conforme el tamaño de la empresa es mayor, representa un menor impacto de movilidad, estando los 4 tamaños de empresas entre los primeros 5 lugares de factores con mayor impacto. Esto probablemente se debe a la adopción del home office por las empresas de mayor tamaño, reduciendo la movilidad de sus empleados en tiempos de dicha pandemia.
- Hallazgo 2: Ahora que tenemos más variables, podemos ver que existen **factores que describen mejor** la situación a comparación de lo que se tenía inicialmente, y esto se ve reflejado con el nivel de significancia de cada variable con el nuevo modelo. A comparación de la **iteración 1**, podemos observar en la que el número de hospitales se redujo un **91.35%** de su valor anterior, pasando a ser la segunda variable menos significativa como atractor principal de viajes, ya que en este nuevo modelo involucramos nuevas variables que explican de mejor manera la cantidad de viajes hacia una comuna.

Identificación de fallas

Algo que se detectó durante la etapa de modelación fue la presencia de overfitting en el modelo, ya que en el proceso no se contaba con alguna inspección relacionada con ello. Este problema fue resuelto antes de finalizar la etapa de modelado, implementando un paso adicional de evaluación del modelo para poder detectar el overfitting.

Revisión de procesos

Diagramas de proceso usando para minería de datos:



El proceso de minería de datos fue similar, los cambios realizados fueron agregar más variables como atractores de viajes mencionados en la evaluación anterior, los cuales son:

- Densidad de la población
- Áreas verdes en metros cuadrados
- Empresas grandes, medianas y pequeñas
- Número de conexiones fijas de internet

Debido a que no se cambió el enfoque de la construcción de datos, no fue necesario construir desde cero el dataset de entrenamiento.

Se tomó el dataset generado durante la última etapa de construcción de datos para únicamente adjuntarle las nuevas variables a considerar.

Revisión de los objetivos de negocio con respecto a sus criterios de éxito.

Recordando que el objetivo es determinar cuáles son los aspectos estructurales que determinan la cantidad de viajes que se realizan hacia una comuna.

Los hallazgos generados por nuestro modelo nos permiten obtener información clara con respecto a cuáles son los principales atractivos de viajes dentro del contexto de nuestro dataset. De momento, contamos con una validación del modelo anterior por parte del socio formador. Este modelo es en esencia el mismo que ya se validó, con la diferencia de que se incluyen más variables en el análisis. Quedamos pendientes a la próxima reunión con el socio formador para validar los resultados de este nuevo modelo.

Siguientes pasos

Opciones a seguir:

Llegados a este punto, contamos con las siguientes posibilidades:

- Definir el modelo como validado y lanzarlo a producción
- Hacer una nueva iteración con cambios que sugiera el SF

Decisión:

En el momento en que se reciba la validación del modelo por parte del SF, el modelo será lanzado a producción. Mientras se espera la validación, se realizarán los preparativos para dicho lanzamiento.