Memoria Descriptiva

El presente proyecto de machine learning nos encontramos enfrente a un ejercicio de clasificación de variables, en donde se intentará determinar dado una serie de características el grado del daño del edificio en caso de un evento sísmico.

El dataset del presente proyecto de machine learning contiene datos relativos al terremoto de Nepal del año 2015. La recolección de los mismo se hace con el propósito de determinar las personas que necesitarían ayudas gubernamentales para la reconstrucción de sus hogares.

Por tanto, el dataset consta solo de edificios cuyo uso principal es la vivienda y algunos con usos secundarios. La encuesta principal consta de una serie de variables que fueron posteriormente purgadas para la competencia, dejando solo aquellas relacionadas con los edificios y el daño sufrido por los mismos.

El target consiste en 3 categorías que denominan el grado daño del edificio, desde daños superficiales que requieren reparaciones cosméticas (categoría 1), hasta daños importantes a la estructura que requerirían la reparación casi integral del inmueble.

Por otro lado, los atributos están constituidos por 38 columnas que describen diversos aspectos de cada edificio evaluado, los cuales podrían dividirse principalmente en 3 categorías: relacionados con el uso del edificio, características descriptivas del edificio y finalmente características con respecto a su ubicación geográfica.

La primera aproximación al modelo fue intentar determinar las variables más importantes en las predicciones, con el objetivo de sólo quedarnos con aquellas que interviniesen verdaderamente en la predicción y mejorar al mismo tiempo la eficiencia computacional.

Para ello, el primer algoritmo implementado fue el ExtraTreeClassifier, que determina la importancia de variables al igual que RandomForest o DecisionTrees, pero al ser aleatorio, es más rápido que estos últimos.

De la feature importance extraída se puede observar que todas las variables relacionadas con la posición geográfica poseen la mayor importancia. Esto tiene sentido, ya que se puede deducir que todas aquellas áreas geográficas cercanas al epicentro del terremoto serían las más afectadas y viceversa. Asimismo la antigüedad del edificio resulta también importante así como la altura. En efecto, la altura de un edificio aumenta su centro de gravedad, lo cual en combinación con los materiales de los cuales está hecho podrían aumentar la propensión a colapsar en caso de un evento sísmico.

Se observa también que todas las categorías que especifican los usos secundarios se ubican en las últimas posiciones tanto como para el RandomForest como para el ExtraTreeClassifier.

Para poder alimentar al algoritmo las categorías que constan de clases codeadas como strings, se intenta convertir dichos atributos con un One Hot Encoding y un Label Encoding, para observar el comportamiento de los algoritmos de predicción para cada uno.

Asimismo, en un intento de mejorar los tiempos de cálculo de los algoritmos, intentamos en primera instancia calcular una predicción simple con todas los atributos presentes, en este caso un RandomForest y un Decision Tree; para luego compararla con los mismo algoritmos sin los atributos que nos parecían redundantes según la Feature Importance mencionada anteriormente.

En este sentido, otro de los algoritmos que intentó fue CatBoostRegressor, el cual en primeras instancias sin modificar parámetros por defecto, arrojaba scores parecidos al Random Forest con el mejor score, a saber 0.72. Sin embargo, una vez los parámetros fueron ligeramente modificados, en especial la profundidad y el numero de iteraciones, el score fue superado por dos puntos, 0.74, y que se comportaba de manera similar en validación cruzada.

Por otro lado, una vez teniendo un score aceptable, se intenta reducir la dimensionalidad del modelo a través de un Label Encoding de las variables geográficas. De esta forma, la variable Geo Level 1 fue reducida de 30 a 16 categorías. Igualmente, se intenta aligerar los tiempos computacionales eliminando las columnas que aparecían en últimos lugares de las diversas feature importances y que además estaban relacionadas con los usos del edificio.

Sin embargo, al hacer la predicción en el nuevo dataset, el score se redujo significativemente, lo cual podría indicar que el Label Encoding fue realizado incorrectamente, así como también que las variables que habían sido eliminadas, en conjunción con otras variables, eran importantes para realizar las predicciones.