Valoración de suelos con «text-mining»

[iStock]/Thinkstock.

La falta de herramientas de valoración inmobiliaria del suelo y la mejora en las técnicas existentes de procesado del lenguaje natural son las bases de este proyecto, cuyo objetivo es predecir el valor de mercado de los suelos a partir de las características encontradas en el campo de descripción de texto libre en anuncios de portales inmobiliarios.

Pedro J. Rodríguez | *Analista de Modelos de Riesgo en Sareb*

En los portales o webs inmobiliarios, es frecuente la existencia de un campo de texto libre en el que el anunciante tiene la posibilidad de detallar mejor las características del inmueble que quiere vender, de forma adicional a lo que permite el resto de campos estructurados del portal. Por otro lado, las herramientas y modelos actuales de valoración existentes en el mercado están muy focalizados en la tipología residencial, siendo poco habituales las aplicables a valoración de suelos.

Por todo ello, decidimos iniciar un proyecto que hará uso de la información contenida en los campos de descripción de texto libre de los anuncios de suelos (urbanos y urbanizables), así como de información externa relevante, para predecir el valor de mercado de dichos activos.

El primer paso antes de organizar las diferentes tareas es obtener un **entendimiento suficiente del negocio inmobiliario** que permita definir y elegir la información más adecuada y relevante para el desarrollo de los modelos. Para ello, tras contactar con varios expertos en valoración inmobiliaria, confirmamos la referencia en la que se sustentan las tasaciones de suelos urbanos y urbanizables realizadas por las entidades tasadoras en el mercado inmobiliario actual, el **método residual de valoración del suelo**, definido en la *Orden ECO/805/2003, de 27 de marzo, sobre normas de valoración de bienes inmuebles*.

De manera resumida, este método considera el valor del suelo como el «*residuo*» que se produce al deducir al valor de venta del producto terminado, todos los costes asociados: gastos de la promoción, costes de construcción y beneficios del promotor. Mediante este método se obtiene el **valor de repercusión del suelo**, medido como precio de venta entre superficie edificable (m² construidos del producto terminado).

Además del método utilizado para valorar, se necesita contar con una información mínima para poder realizar la valoración de un suelo: la **superficie total del terreno**, la **localización** (municipio y provincia en el que se encuentra el suelo, distancia al centro urbano, etc.), la **clasificación o tipología del suelo** (suelo urbano, urbanizable y no urbanizable o rústico), la **calificación o usos permitidos del suelo** (residencial, terciario, industrial, dotacional, etc.) y la **edificabilidad** o superficie máxima edificable.

La metodología seguida para afrontar el problema propuesto ha sido la metodología **CRISP-DM** (*Cross-industry standard process for data mining*), que define el marco de trabajo a seguir para el correcto desarrollo de un proyecto de *Data Science*.

Una vez conocemos mejor el negocio, organizamos las tareas del proyecto, y definimos la metodología a aplicar, se divide el problema en cuatro fases:

Fase 2: Desarrollo del modelo de «text-mining»

El objetivo de esta fase es obtener el valor de las características de los suelos dentro del campo de texto libre, mediante el desarrollo de un modelo de extracción de información o modelo de reconocimiento de entidades (NER, *Named Entity Recognition*). Se decide utilizar modelos de *Deep Learning*, utilizando una muestra de datos de entrenamiento etiquetados a mano. El mejor modelo resultante es una RNN (red neuronal recurrente) con tres capas LSTM (*Long Short-Term Memory*) establecidas secuencialmente, añadiendo un 30% de *dropout* no recurrente. Este modelo mejora en un 28% el rendimiento (métrica *F-Score*) del *modelo benchmark* propuesto, definido como aquél que clasifica todas las palabras del texto como «sin significado», como queda confirmado en la evolución del *accuracy* de cada modelo según aumenta el número de épocas entrenadas.

FASES DEL PROYECTO DE VALORACIÓN DE SUELOS CON «TEXT-MINING»

Fuente: *elaboración propia*.

Fase 1: Obtención de datos

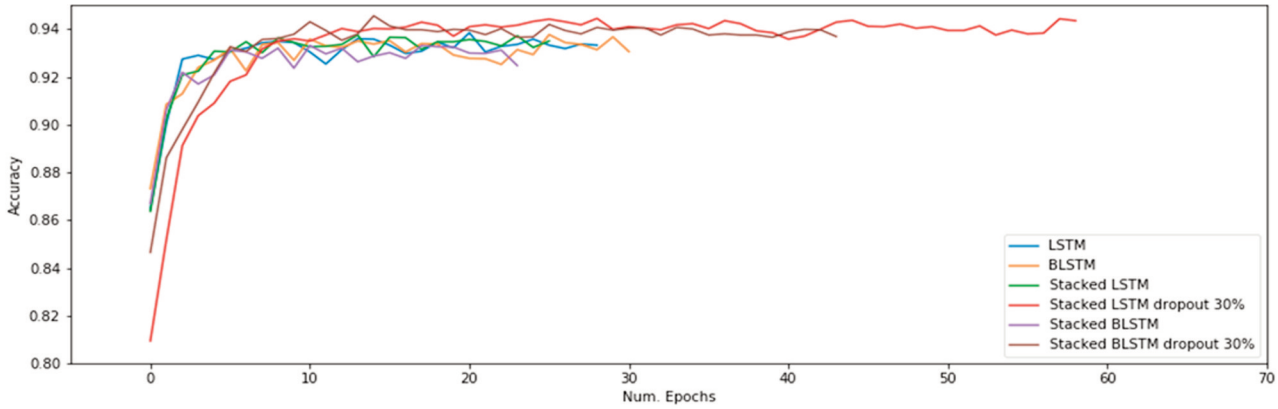
Se procede a la descarga de un mínimo número de anuncios con datos suficientes, utilizando técnicas de web scrapping. Se extraen datos tanto estructurados (en forma de «clave-valor») como no estructurados (descripción de texto libre), y se almacenan en una base de datos NoSQL -en nuestro caso MongoDB. Analizamos y tratamos estos datos mediante técnicas de NLP (procesamiento de lenguaje natural) y expresiones regulares.

Se desarrollan, además, un **modelo de extracción de la tipología de suelo** cuyo objetivo es la extracción de la tipología del suelo para la selección de la muestra de registros a etiquetar para el modelo NER, y un **modelo de extracción de la localización** para la obtención de la provincia y municipio a partir de la localización escrita en texto. Ambos modelos utilizan, fundamentalmente, expresiones regulares.

Fase 3: Desarrollo del modelo de precios

El objetivo, en esta fase del proyecto, es estimar el

Evolución del *accuracy* en validación durante *training* (tamaño ventana 3)



Fuente: *elaboración propia*.

precio final del inmueble anunciado. Utilizamos los datos extraídos del campo de descripción de texto libre (predicciones del modelo de reconocimiento de entidades de *text-mining*) y datos incorporados de fuentes externas como el INE (Instituto Nacional de Estadística), IGN (Instituto Geográfico Nacional), Fomento, Idealista y CYPE (costes de construcción), entre otros, sin utilizar ningún dato adicional disponible en los anuncios.

Se entrenan diversos modelos de regresión incluidos en la librería *scikit-learn* de Python, desde los más sencillos, como regresiones lineales o árboles de decisión, hasta modelos de *ensemble* más complejos, como *Gradient Boosting* y *Random Forest*.

El mejor modelo de regresión obtenido en esta fase es un *Random Forest* (*ensemble* de árboles de decisión), cuyo rendimiento, medido en términos de R^2 ajustado, resulta en una mejora de un 121% con respecto al *modelo benchmark* definido, el cual utiliza únicamente los datos estructurados de los anuncios.

Se aprecia que las variables con más relevancia en el modelo final tienen **sentido de negocio**, como la edificabilidad, la superficie total o la distancia al núcleo urbano.

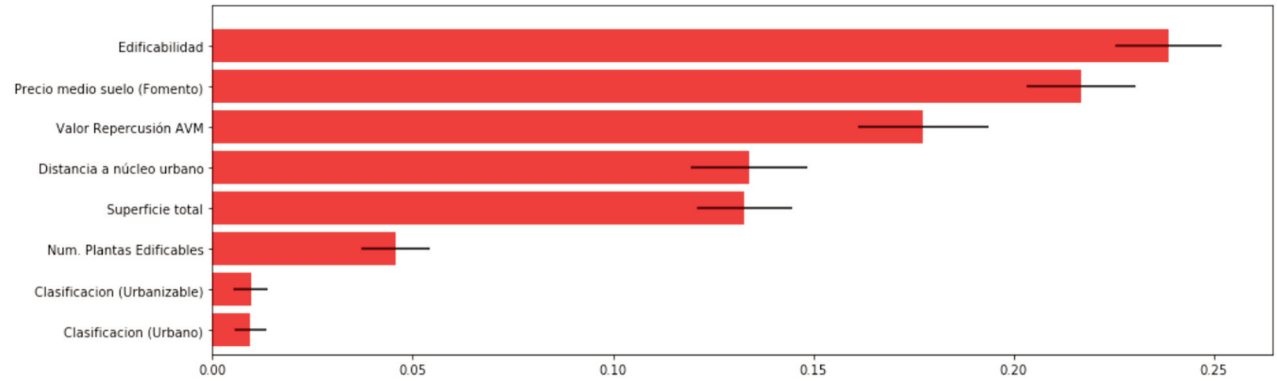
Adicionalmente, se desarrolla un modelo sencillo de valoración estadística (AVM), que implementa el **método residual estático** (*artículos 40 a 42 de la Orden ECO/805/2003*) de valoración inmobiliaria. Toma como datos de entrada el tipo de suelo y la lista de usos permitidos, y devuelve el **valor de la repercusión del suelo**. Este valor se añade también como *input* al modelo de regresión de predicción del precio.

Fase 4: Integración final

En la última fase, integramos todos los pasos y resultados intermedios para generar un **proceso único e integrado**, mediante el cual, a partir del texto con las características de un suelo, se pueda estimar su valor de mercado. De esta manera, facilitamos su uso en posibles aplicaciones futuras. Finalmente, como muestra de aplicación práctica del proyecto, se desarrolla una aplicación web sencilla que utiliza, de manera integrada, todos los modelos descritos.

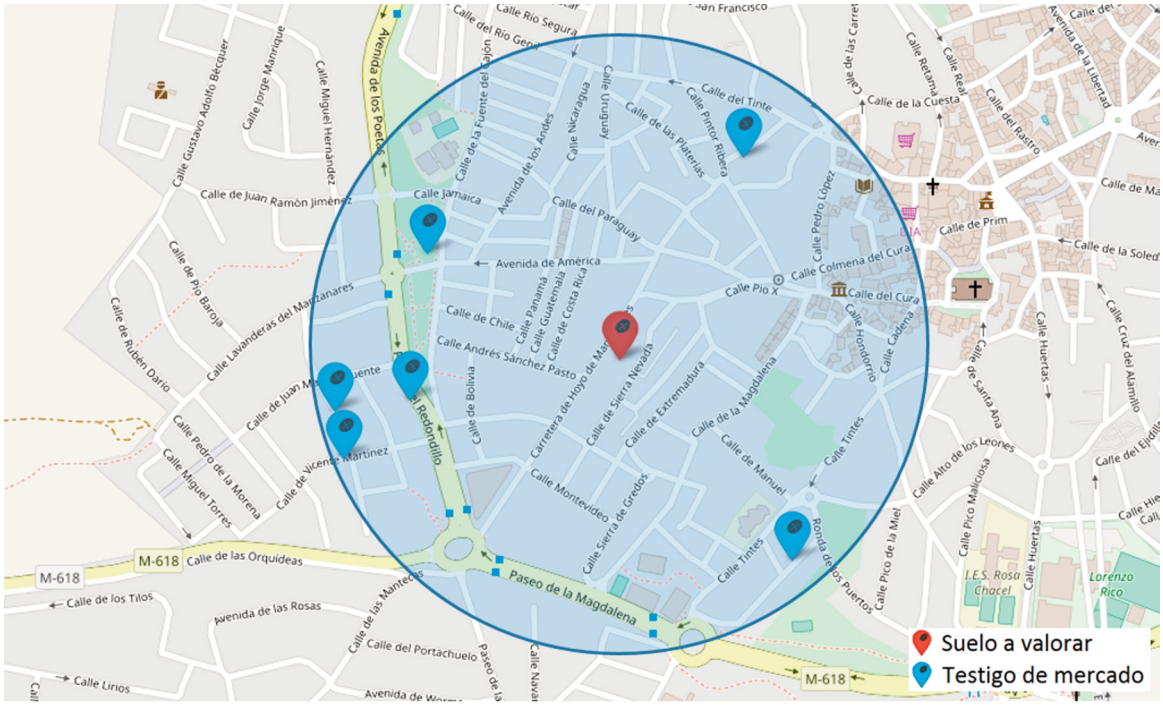
En esta aplicación, el usuario únicamente tiene que introducir la descripción y la localización del

Importancia de las variables



Fuente: *elaboración propia*.

EJEMPLO GRÁFICO DE TESTIGOS*



*Ejemplo gráfico de selección de testigos de mercado (comparables) para el cálculo del AVM.

Fuente: *elaboración propia*.

APLICACIÓN WEB

Fuente: *elaboración propia*.

suelo a valorar en lenguaje natural, y mediante un único *click* obtendrá el valor de mercado de dicho suelo estimado por el modelo ::

Nota: este artículo es un extracto del trabajo de fin de curso del Máster Executive en Data Science y Big Data en Finanzas 2017-2018, Afi Escuela de Finanzas.