Segunda Entrega del Proyecto

Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Juan José Toro Villegas-Julián Vargas Uribe-Arthur Jose Mosquera Franco

El proyecto se enfoca en el conjunto de datos de Zillow Prize, que contiene información sobre las transacciones de bienes raíces y los valores de las propiedades en diferentes partes de los Estados Unidos. El objetivo principal del proyecto es utilizar técnicas de análisis de datos para explorar las relaciones entre las variables y predecir los valores de las propiedades, en este caso con el Log Error de estas.

Avances

En primer lugar, la importación de los datos. Siendo estos inicialmente 4: Train de los años 2016 y 2017 (2); y properties 2016 y 2017 (2). Después se concatenaron los archivos para juntar toda la información y así solamente quedarnos con dos datasets, todo lo respectivo al 2016 e igual con el 2017.

Exploración Inicial

Se realizó una exploración inicial de los datos y se identificó la estructura general del conjunto de datos y sus variables: conteo de datos, distribuciones, promedios, correlaciones con la variable objetivo, histogramas de cada variable, desviaciones estándar, valores mínimos y máximos, tipo de datos de cada variable, entre otros.

La variable objetivo "Logerror" vemos que se distribuye normalmente, con un sesgo bajo de 3,70.

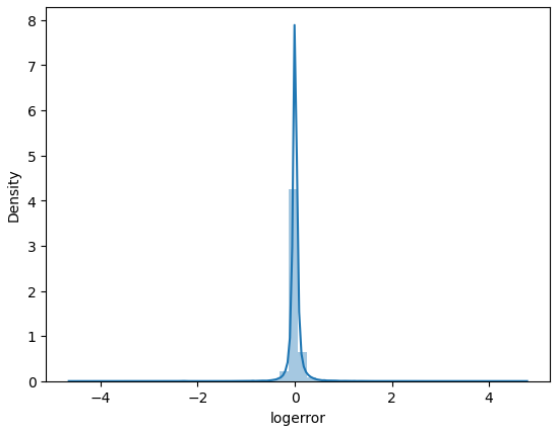

imagen 1.

En cuanto a los datos faltantes, esto significó uno de los grandes desafíos ya que muchas variable mostraron un alto porcentaje de datos faltantes, algunos fueron:

basementsqft	1282	67.723191
buildingclasstypeid	1282	67.723191
finishedsquarefeet13	1282	67.723191
fireplaceflag	1281	67.670365
architecturalstyletypeid	1280	67.617538
typeconstructiontypeid	1280	67.617538
finishedsquarefeet6	1279	67.564712
pooltypeid10	1276	67.406233
decktypeid	1272	67.194929
poolsizesum	1268	66.983624
pooltypeid2	1260	66.561014
hashottuborspa	1254	66.244057
finishedsquarefeet15	1235	65.240359
taxdelinquencyyear	1227	64.817750
taxdelinquencyflag	1227	64.817750

imagen 2.

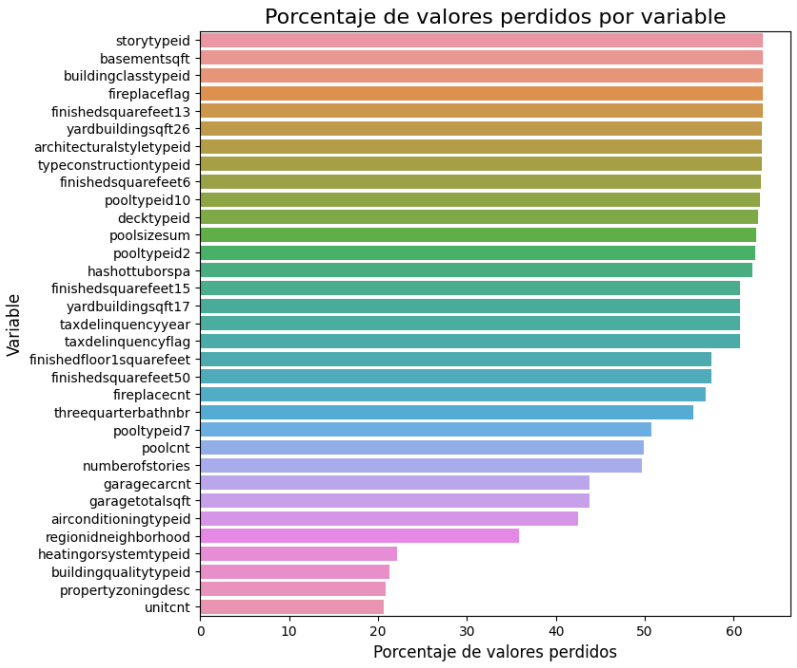

imagen 3.

por tanto se decidió que se depurarian las variables es decir, columnas que presentaron un porcentaje de datos faltantes mayor a 70%, en conclusión se eliminaron casi el 60% de columnas iniciales.

Posteriormente se llenaron los datos faltantes con el promedio de la respectiva columna.

También nos percatamos de que la variable objetivo no tuviese valores que tienen lectura cero, ya que esto puede afectar la predicción del modelo más adelante.

Añadir también que las variables tipo “object” es decir String, le aplicamos una transformación para así estas puedan darnos más información a la hora de analizarlas, siendo así entonces luego de ver y transformar algunos datos iniciales quedamos con types: float64(25), int64(9), en referencia a los tipos de variables. (solo corresponde al dataset año 2017).

Otro inconveniente que presentamos fue que para el año 2016, no hay mucha información, es decir que tal dataset creemos solo brinda muy poca información, posee muchos datos faltantes, mucho más que el año 2017.

Por último decir que para toda la parte del análisis se crearon gráficos y visualizaciones para ayudar a comprender mejor los patrones y relaciones que los datos presentan y que todo esto será analizado mucho más a profundidad para la entrega final del proyecto junto con el modelo propuesto por nosotros para cumplir con la competencia seleccionada.

Enlace de la competencia de Kaggle:
https://www.kaggle.com/competitions/zillow-prize-1/data?select=zillow_data_dictionary.xlsx

Enlace a los datasets en Drive:
https://drive.google.com/drive/folders/1lGXke4GmGdNLxM31L24A-y5NQydNJrXb?usp=share_link