Proyecto Individual 2- Data 04- Soy Henry

Data Engineering

En la siguiente práctica número 2 tenemos un cliente que nos da 2 dataframe con información de precios de inmuebles en Colombia,

nos recalca que lo que quiere predecir es que si un inmueble es caro o barato, y que esta condición la obtengamos a partir del promedio de precios de todo el dataset.

Se debe usar solo uno de los datasets "propiedadesTrain" para hacer la transformación segun nuestro propio criterio y así dejar el dataset lo mas limpio posible y con las variables que a nuestro juicio sean las mas influyentes para poder predecir un valor de un inmueble, desafortunadamente estos datos como es normal tiene muchas deficiencias, ya que hay muchas varibles que carecen de información muy valiosa para que nuestro trabajo sea un poco más fácil, pero esta es la realidad con la que nosostros como data Engineering nos encontramos a diario, así que debemos encontrar siempre la mejor forma para poder transformar de la mejor manera este datasets y que nuestro dataset final sea óptimo para dar así buenos resultados.

En nuestro proceso de ETL nos encontramos con muchas situaciones en las que debemos tomar desiciones muy difíciles con rspecto a si es mejor borrar o remplazar valores nulos o faltantes, yo analicé una situación en particular, con las columnas que nos dan la fecha de alta de un aviso y de baja, y saqué como conclusión que si un inmueble al ponerse a la venta no demora mucho en venderse, quiere decir que este precio del inmueble es muy bueno y accesible para cualquier persona interesada en comprar, también teniendo en cuenta la ubicación real del inmueble, por eso deje para ubicar bien los inmuebles solo la columna de longitud y latitud.

Esta práctica en conclusión es muy buena para ir viendo el contexto de este problema en particular y así compararlo con situaciones parecidas, y me doy cuenta definitivamente lo indispensable que es tener una buena data, y que al no tenerla debemos usar mucho de nuestro ingenio, usando herramientas y librerias que nos ayuden para no perder información importante de dichos datos.

Tambien lo importante de poder realizar todo el proceso de ETL lo mejor posible y tener tiempo de probar con varios modelos y así escoger el de mejor resultados...

Logro hacer dos modelos de predicción uno con regresión logística y otro con random forest, este último me dio mejores resultados en las métricas.

Así que sin duda escojo este último para que me evalúen la práctica.

no es mas ppor ahora, gracias