Primera Entrega del Proyecto

Por:

Juan José Toro Villegas
Julián Vargas Uribe
Arthur Jose Mosquera Franco

Materia:

Introducción a la Inteligencia Artificial para las Ingenierías

Profesor:

Raúl Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín 2023

Descripción del Problema

Zillow es una compañía tecnológica que opera como un mercado para bienes raíces en línea. La compañía fue fundada en el 2006 por Rich Barton, Lloyd Frink y Spencer Rascoff en Estados Unidos, y tiene como modelo de negocio la venta de anuncios de viviendas en su sitio web y ahora en su aplicativo. Allí puedes tanto comprar, rentar y vender propiedades. (Gutiérrez, 2021. Con información de Wikipedia y Motley Fool).

Los "Zestimates" son valores estimados de viviendas basados en 7,5 millones de modelos estadísticos y de aprendizaje automático que analizan cientos de puntos de datos en cada propiedad. Y, al mejorar continuamente el margen de error promedio (del 14 % al principio al 5 % en la actualidad), Zillow se ha establecido desde entonces como uno de los mercados más grandes y confiables para la información de bienes raíces en los EE. UU. Por tanto, es indispensable mejorar el cálculo y ese margen de error promedio para así contar con una mayor exactitud y en menor tiempo a la hora de hablar de valor de bienes raíces. (Andrew Martin, Bin, Cat N, K Nielsen, Maggie, Wendy Kan, Zillow Prize: Zillow's Home Value Prediction (Zestimate) publicado en Kaggle, 2017).

Objetivo

Desarrollar un algoritmo que haga predicciones sobre los precios de venta futuro de las viviendas, apoyadas en datos de transacciones de bienes raíces que son información pública del año 2016 y 2017.

Dataset

Vamos a usar el dataset de Kaggle de esta competición (https://www.kaggle.com/competitions/zillow-prize-1/data), que tiene 1,048,576 número de muestras y las siguientes columnas:

Feature	Description
'airconditioningtypeid'	Type of cooling system present in the home (if any)
'architecturalstyletypeid'	Architectural style of the home (i.e. ranch, colonial, split-level, etc…)
'basementsqft'	Finished living area below or partially below ground level
'bathroomcnt'	Number of bathrooms in home including fractional bathrooms
'bedroomcnt'	Number of bedrooms in home
'buildingqualitytypeid'	Overall assessment of condition of the building from best (lowest) to worst (highest)
'buildingclasstypeid'	The building framing type (steel frame, wood frame, concrete/brick)
'calculatedbathnbr'	Number of bathrooms in home including fractional bathroom
'decktypeid'	Type of deck (if any) present on parcel

'threequarterbathnbr'	Number of 3/4 bathrooms in house (shower + sink + toilet)
'finishedfloor1squarefeet'	Size of the finished living area on the first (entry) floor of the home
'calculatedfinishedsquarefeet'	Calculated total finished living area of the home
'finishedsquarefeet6'	Base unfinished and finished area
'finishedsquarefeet12'	Finished living area
'finishedsquarefeet13'	Perimeter living area
'finishedsquarefeet15'	Total area
'finishedsquarefeet50'	Size of the finished living area on the first (entry) floor of the home
'fips'	Federal Information Processing Standard code - see https://en.wikipedia.org/wiki/FIPS_county_code for more details
'fireplacecnt'	Number of fireplaces in a home (if any)
'fireplaceflag'	Is a fireplace present in this home
'fullbathcnt'	Number of full bathrooms (sink, shower + bathtub, and toilet) present in home
'garagecarcnt'	Total number of garages on the lot including an attached garage
'garagetotalsqft'	Total number of square feet of all garages on lot including an attached garage
'hashottuborspa'	Does the home have a hot tub or spa
'heatingorsystemtypeid'	Type of home heating system
'latitude'	Latitude of the middle of the parcel multiplied by 10e6
'longitude'	Longitude of the middle of the parcel multiplied by 10e6
'lotsizesquarefeet'	Area of the lot in square feet
'numberofstories'	Number of stories or levels the home has
'parcelid'	Unique identifier for parcels (lots)
'poolcnt'	Number of pools on the lot (if any)
'poolsizesum'	Total square footage of all pools on property
'pooltypeid10'	Spa or Hot Tub
'pooltypeid2'	Pool with Spa/Hot Tub
'pooltypeid7'	Pool without hot tub
'propertycountylandusecode'	County land use code i.e. it's zoning at the county level
'propertylandusetypeid'	Type of land use the property is zoned for
'propertyzoningdesc'	Description of the allowed land uses (zoning) for that property
'rawcensustractandblock'	Census tract and block ID combined - also contains blockgroup assignment by extension
'censustractandblock'	Census tract and block ID combined - also contains blockgroup assignment by extension
'regionidcounty'	County in which the property is located
'regionidcity'	City in which the property is located (if any)
'regionidzip'	Zip code in which the property is located
'regionidneighborhood'	Neighborhood in which the property is located
'roomcnt'	Total number of rooms in the principal residence

'storytypeid'	Type of floors in a multi-story house (i.e. basement and main level, split-level, attic, etc.). See tab for details.
'typeconstructiontypeid'	What type of construction material was used to construct the home
'unitcnt'	Number of units the structure is built into (i.e. 2 = duplex, 3 = triplex, etc...)
'yardbuildingsqft17'	Patio in yard
'yardbuildingsqft26'	Storage shed/building in yard
'yearbuilt'	The Year the principal residence was built
'taxvaluedollarcnt'	The total tax assessed value of the parcel
'structuretaxvaluedollarcnt'	The assessed value of the built structure on the parcel
'landtaxvaluedollarcnt'	The assessed value of the land area of the parcel
'taxamount'	The total property tax assessed for that assessment year
'assessmentyear'	The year of the property tax assessment
'taxdelinquencyflag'	Property taxes for this parcel are past due as of 2015
'taxdelinquencyyear'	Year for which the unpaid propert taxes were due

Métricas

Como métrica de Machine Learning usaremos el MAE (Mean Absolute Error) entre el error de registro previsto y el error de registro real. Esto para facilitar y cuantificar la precisión del modelo, el cual se espera que tenga un porcentaje de acierto alto y que a su vez se vea reflejado en la cantidad de personas que usan la aplicación a la hora de hablar de bienes raíces.

El error de registro definido así:

$$log_{error} = log_{Zestimate} - log_{precio\ venta}$$

Como métrica de negocio, gracias a la utilización del modelo, esperamos un incremento de al menos un 35% en ventas ya que, debido a las predicciones acertadas de la aplicación y pagina web, los clientes gozaran de un nivel de confianza mayor, a la hora de usar e invertir en Zillow.

Primer Criterio

Lo que esperamos de la implementación del modelo es que este de cifras muy cercanas a la realidad del precio de las viviendas, que los americanos puedan apoyarse también del modelo a la hora de tomar una decisión de comprar, rentar o vender una propiedad y es realmente lo que se quiere al usar Zillow.

La variable que más valor tiene para una organización son las ventas (que tanto incrementan); sin embargo, para un primer momento y como decisión tomada de la junta directiva, la variable que inicialmente evaluará al modelo en los primeros meses será el número de visitas mensuales tanto del aplicativo como de la página web en las ciudades más importantes de

EE. UU. Esto no quiere decir que se dejaran de lado el crecimiento de ventas planteado, solo que pasa a un segundo plano en los primeros meses de implementación.

NOTA: Los criterios están sujetos a cambios, a medida que se desarrolla el modelo.

Bibliografías.

https://www.kaggle.com/competitions/zillow-prize-1/data?select=zillow_data_dictionary.xlsx