Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales
Facultad de Ingeniería

Mayo 2022

Aprendizaje Automático

Trabajo Práctico N° 1

Grupo N°3

Altschuler, Florencia ; Krell, Federico ; Picasso, Juan Pablo

Resumen

El objetivo de este trabajo fue analizar el conjunto de datos de precios de propiedades en Capital Federal desde el punto de vista del aprendizaje automático, la importancia de las variables obtenidas y, además, profundizar en el conocimiento de los atributos de árboles de decisión. Se obtuvieron los modelos optimizados para dos métricas y a partir de ello se llegó a la conclusión que la variable de mayor interés para definir el precio es la superficie cubierta y si pertenece o no a la zona norte de la ciudad.

Introducción

Properati es un portal web de propiedades sobre avisos de inmuebles de toda Latinoamérica desde 2015. En el presente trabajo se analizaron los inmuebles en venta de Capital Federal con el objetivo de predecir el precio categorizado como Bajo, Medio y Alto con una distribución de forma balanceada en cada categoría.

Se utilizó un modelo de Árbol de clasificación de la librería de Sci-Kit Learn.

Inicialmente, se presentan los datos utilizados y se describen las variables seleccionadas para las etapas posteriores. Luego, se desarrollan las técnicas y los criterios utilizados para la limpieza y análisis de los atributos de la base de datos. En esta sección también se describe el modelo de aprendizaje automático supervisado junto con los atributos y parámetros seleccionados.

Por último, se desarrolla el refinamiento de hiper parámetros utilizando el algoritmo de búsqueda aleatorizada con validación cruzada en 5 partes.

Datos

La base de datos utilizada contenía 183810 observaciones y 26 atributos con datos de distintos tipos. Los datos geográficos (i.e. 'lat' y 'lon'), temporales (i.e. 'start_date', 'end_date' y 'created_on'), no estructurados (i.e. 'title' y 'description') y las columnas 'Unnamed: 0', 'Unnamed: 0.1' no se utilizaron. Las columnas 'l1', 'l2', 'l5', 'l6', 'ad_type' y 'operation_type' se quitaron porque eran constantes. Las columnas 'l4' (con los valores: 'Palermo Hollywood', 'Palermo Soho', 'Palermo Chico','Palermo Viejo') y 'price_period' se quitaron porque había muchos registros con valores faltantes. Además, se conservaron solo aquellas propiedades cuyo precio estuviese expresado en dólares (i.e. 'currency' = 'USD').

Nombre de la variable	Tipo de variable	Medida de tendencia central	Rango de valores	Datos faltantes (NA's)
`'rooms'`	Cuantitativa discreta	2.706719 (media)	1 a 40	168242
`'bedrooms'`	Cuantitativa discreta	1.983696 (media)	0 a 153	49976
`'bathrooms'`	Cuantitativa	1.584221 (media)	1 a 20	24918

Nombre de la variable	Tipo de variable	Medida de tendencia central	Rango de valores	Datos faltantes (NA's)
	discreta			
'price'	Cuantitativas continuas	282119 (media)	4300 a 35000000	3392
'surface_total'	Cuantitativas continuas	174.63733 (media)	10 a 140380	60902
'surface_covered'	Cuantitativas continuas	148.706692 (media)	1 a 950000	62528
'l3'	Cualitativa nominal	"Palermo" (moda)	-	1847
'property_type'	Cualitativa nominal	"Departamento" (moda)	-	0

Tabla 1. Descripción de las variables utilizadas en el análisis.

Los registros con algún dato faltante de precio, superficie total o superficie cubierta, no se consideraron en el análisis. El resto de los datos faltantes se imputaron con algún estadístico de tendencia central (la media para las variables cuantitativas y la moda para las cualitativas nominales).

Los registros con algún valor atípico (ej. precios por debajo de USD 1000 o números de habitaciones negativos) se eliminaron, representando en total 5 registros.

Metodología

Todo el análisis se realizó en *Python* 3.7 y se utilizaron las librerías *Numpy, Pandas* y *Sklearn.* Algunos de los módulos de la librería Scikit-Learn que se emplearon fueron *model_selection, ensemble* y *metrics*.

Se dividió al conjunto de datos por terciles en función del precio `'bajo'`, `'medio'` y `'alto'` de forma que las clases queden balanceadas.

Luego de la distribución se procedió con la limpieza del dataset. En la limpieza se identificaron las columnas que eran relevantes y aquellas que podrían no utilizarse como se explica en la sección datos. Posteriormente, se hizo una limpieza de los datos que no tenían información sobre superficie. Con el objetivo de simplificar el modelo se decidió no imputar los valores que por experiencia sabíamos que iban a ser sensibles.

Por otro lado, aquellas propiedades que no tenían precios también se eliminaron ya que no podríamos hacer predicciones imputando valores de la columna objetivo.

Se obtuvo el árbol que denominamos como "sencillo" y luego, mediante el algoritmo de *Randomized Search*, se buscaron los mejores parámetros para un árbol con métrica F1. Las combinaciones de parámetros utilizadas se presentan en la tabla 2.

	Criterio	Profundidad	Mínimo de muestras por hoja	Parámetro de poda
Árbol sencillo	Gini	4	-	-
Árbol F1	Gini	42	1	0

Tabla 2. Parámetros utilizados en cada modelo de clasificación.

Resultados

Las variables más importantes (feature importance) a la hora de clasificar los precios resultaron ser la superficie cubierta, superficie total, la latitud y si la propiedad se ubica en el barrio de Puerto Madero.

Se puede ver en el siguiente gráfico el Árbol sencillo generado:

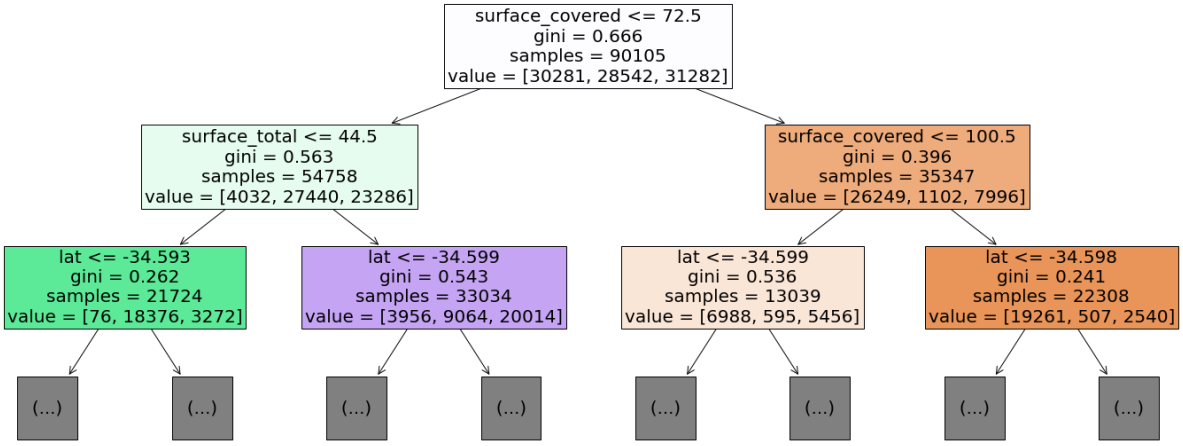

Figura 1. Árbol sencillo (Simple tree). Se acotó la cantidad de nodos para mejorar su presentación. Referencias: color naranja categoría precio alto (color naranja), categoría precio medio (color violeta) y categoría precio bajo (color verde).

En la siguiente tabla se muestra la importancia de cada variable para los dos árboles.

Variable	Árbol sencillo	Árbol complejo
surface_covered	0.584	0.354
surface_total	0.314	0.226
lat	0.096	0.173
lon	-	0.127
rooms	-	0.019
bedrooms	-	0.016
l3_Puerto Madero	0.0065	0.013

Tabla 3. Importancia de las variables según cada tipo de árbol de decisión.

Estos resultados fueron coherentes con lo esperado ya que la latitud que divide las propiedades es -34.5, que corresponde a la Av. Corrientes, la cual divide el corredor Norte (i.e. comunas 1, 2, 14 y 13) y el resto de la ciudad.

Por último, las variables más importantes se mantuvieron en ambos árboles, obteniéndose que en el segundo si bien se agregaron factores los originales mantienen su volumen inicial.

Conclusiones

Se llegó a ciertas conclusiones de las diferentes experiencias del trabajo. La primera fue que el árbol se separó efectivamente como se esperaba en base al conocimiento popular. Esto se observó claramente en la división basada en la variable latitud y en el mayor precio para la propiedades del barrio de Puerto Madero.

Por otro lado, pudimos ver la mejora que se obtuvo al desarrollar una búsqueda de un árbol optimizado mediante las herramientas de *random search*.

Por último, se decidió el uso del *F1 score* balanceado como métrica ya que, teniendo en cuenta que el objetivo es la optimización general de la matriz de confusión, nos proporciona una respuesta equilibrada entre *accuracy* y *recall*.

En función al análisis desarrollado, se sugiere que para la compra de un inmueble en este distrito, la ubicación es el aspecto de mayor peso. En los casos de un presupuesto más acotado, se deberían considerar propiedades ubicadas en las latitudes más bajas de la ciudad. Otras variables que podrían aportar a futuros análisis son, por ejemplo, la distancia a ciertos medios de transporte públicos (e.g. metrobus y subterráneos) y la ubicación de zonas recreativas y/o espacios verdes.

Bibliografía

- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. Link: https://scikit-learn.org/
- Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. Aurélien Géron, O'Reilly, 2da edición, 2019.
- Material de clase.