Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones

Informe Trabajo Práctico 3 Mentoría:

"Análisis del mercado inmobiliario de la Ciudad de Buenos Aires 2017"

Mentor: Javier Lezama

Integrantes Grupo:

-Navarro Agustina

-Calle Manuel

-Tagle Gabriel

Julio 2019

Introducción

Para la confección del informe se desarrolló una notebook en Jupyter donde se aplicaron los conocimientos de la primera materia de la diplomatura: “Introducción al Aprendizaje Automático”.

En la notebook usamos diferentes librerías que nos permitieron combinar herramientas de manejo de grandes volúmenes de datos, análisis estadísticos, división de datos, y algoritmos de predicción.

El mentor de este grupo es Javier Lezama quien nos orientó en lo relacionado al contenido del dataset y a la descripción del caso para poder desarrollar este informe.

División de datos en conjuntos de entrenamiento y evaluación.

Para comenzar el análisis, se plantearon las variables que son importantes para poder estimar la variable precio de la propiedad en dólares. Se definieron las siguientes:

- Superficie Total en m2 ('surface_total_in_m2')
- Codigo de Barrio ('barrio_cod')
- Tipo de Propiedad ('property_type')
- Cantidad de Habitaciones ('rooms')
- Superficie Cubierta en m2 ('surface_covered_in_m2')

Aplicando “train_test_split” de la librería “sklearn.model_selection” se dividió el conjunto de datos en 2 grupos. Uno de validación y uno de entrenamiento, con el 20% y 80% del total del dataset respectivamente.

Elección de un modelo de Regresión

El modelo de regresión elegido fue Polinomial de grado 4 (cuatro) sin término de regularización. Se eligió este modelo luego de aplicar la técnica GridSearchCV para regresión lineal, polinomial con y sin regularización tanto Lasso como Ridge. Siendo el modelo de regresión polinomial grado 4 sin regularización el que arrojó menor error cuadrático medio se optó por este modelo. En el siguiente punto del práctico se detalla este proceso.

Selección de Hiperparámetros

Cómo se mencionó previamente, para la selección de hiper parámetros se aplicó la técnica de GridSearchCV, que permite hacer una búsqueda exhaustiva de varios parámetros para obtener el mejor resultado de un modelo de aprendizaje automático. Se definió como función de costo la del error medio cuadrático y se buscó minimizar este valor en cada caso. En todos los casos, se definió el parámetro cv=5 para poder aplicar cross validation en 5 bloques para el conjunto de datos de entrenamiento.

Entonces, se planteó un GridSearchCV para regresión polinomial con el parámetro del grado definido desde 1 al grado 11. El mejor resultado en este caso fue el de grado 4 con MSE = 949,188,925.35

A su vez, se planteó el mismo proceso para regresión polinomial con regularización Lasso y Ridge pero no se llegó a mejores resultados por lo que se continuó el estudio con el modelo definido anteriormente. Se aclara que la elección del MSE puede verse distorsionada por la presencia de outliers ya que, al elevar los residuos al cuadrado, se potencia el efecto de los outliers. Por lo que se podría haber optado por calcular RMSE o algún otro cálculo de los errores.

Métricas sobre el conjunto de evaluación

Finalmente, se realizó la predicción sobre el conjunto de validación, aplicando el modelo entrenado anteriormente, obteniendo un MSE menor al obtenido en el conjunto de entrenamiento:

- MSE Validación = 750,345,218.14

La métrica definida fue R2 Score y se obtuvo un valor de 0.97, lo cual muestra que las predicciones con este modelo tendrán un 97% de efectividad.

Conclusiones

Este trabajo práctico permitió al grupo aplicar conocimientos necesarios para entender el concepto del aprendizaje automático y la elección de un modelo que permita predecir los precios de las propiedades buscando el mayor grado de exactitud.

En este práctico se exploraron modelos de regresión lineal y polinomial, para la elección del mejor modelo y luego analizar los hiperparámetros para definir los que minimizan el error cuadrático medio. La elección de un modelo de regresión para predecir datos, puede ser una forma de encontrar rápidamente un modelo simple que pueda predecir con exactitud nuestra variable objetivo del set de datos.

Sin embargo, aún se puede seguir indagando en mejorar la predicción pensando en desafiar más algunos parámetros como, validar si fueron acertadas las variables del set de datos utilizadas o limpieza de outliers antes de aplicar el modelo elegido.

A lo largo de los siguientes prácticos se irán aplicando nuevos conceptos hasta finalmente optimizar el modelo de predicción para poder obtener la mayor precisión con el menor costo computacional.