Modelos y resultados

William Aguirre & John Esteban Londoño

2024-05-26

Modelos y resultados

Para entrenar los modelos se utilizó la técnica de validación cruzada, de manera que se pudiera reducir el efecto de la autocorrelación espacial sobre el aprendizaje de cada modelo. Teniendo en cuenta que el Marco Geoestadístico Nacional del DANE provee una agrupación geográfica a nivel de manzana, sección (conjuntos de manzanas) y sectores (conjuntos de secciones), se pudo utilizar cada sección como un fold que rompiera con la dependencia espacial del precio.

```
sectores<-unique(train$cod_sector)
folds<-createFolds(sectores,k=length(sectores),list = TRUE,returnTrain = TRUE)
```

Los modelos se entrenaron en primera instancia con la información extraida de la base de datos y del censo del DANE `nbanios`, `nhabitaciones`, `piso_apartamento` ; `estrato` . Dado que los precios de las viviendas también corresponden a diferentes meses de un horizonte de tiempo de tres años, se tuvo que incluir la variable `periodo` para tratar de establecer un efecto fijo del tiempo sobre el precio. Posteriormente se fueron incluyendo las variables de seguridad y salud, encontrando que hubo una mejora significativa en la predicción al incluir este tipo de variables.

Los algoritmos utilizados para las predicciones fueron *Regresión lineal*[*], *Elastic net* [**], *Random forest* [***]y *Xgboost* [****]. Cada uno de estos se entrenó realizando la validación cruzada con los folds anteriormente señalados que pueden ser observados en Mapa de folds.

Conclusiones y recomendaciones

Se logra evidenciar que las caracteristicas de una vivienda son un buen predictor de su precio, sin embargo, las variables del entorno también se convierten en un factor importante al momento de establecer su precio. Evidenciamos como hay aspectos del entorno que pueden afectar positivamente su precio, como la distancia a amenities como hospítales, mientras que existen otras variables que pueden impactar negativamente el precio, como el aumento de la criminalidad en la zona.

También observamos que en este tipo de problemas algoritmos que tienen un poca carga computacional como Elastic Net y la propia regresión lineal tienen un rendimiento aceptable frente a algunos modelos más complejos como Xgboost. Por lo que el tiempo dedicado a entrenar modelos podría bien utilizarse en mejorar la modelación en una regresión y obtener resultados mucho mejores y menos costosos.

La cantidad de variables del entorno incluidas impacta significativamente el rendimiento de los modelos, por lo que se recomienda en estos problemas dedicar un mayor tiempo a la construcción de un set de datos lo más completo posible. En este sentido, por restricciones de tiempo no fue posible incluir otras variables que podrían mejorar la predicción como la distancia a centros comerciales y zonas verdes.

Table 1: Resumen de resultados

Algoritmo	Variable objetivo	Predictores utilizados	MAE fuera de muestra
Regresión lineal	price	nbanios, nhabitaciones, piso_apartamento, estrato	860024814
Regresión lineal	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato	295197752
Regresión lineal	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo	294830576
Elastic net	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo	283134955
Elastic net	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	268696039
Elastic net	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	276791195
Elastic net (caret)	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	264342673
Random forest	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo	274216139
Random forest	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	244889109
Random forest	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	252533214
Random forest (caret)	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	249855878
Random forest	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia, distancia_hospital	235325152
Xgboost	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	277450159