anexoPLSDA

Ana Valiente

2025-06-06

ANEXO PLS DISCRIMINANTE

Con el objetivo de predecir las **fluctuaciones del precio** en las viviendas valencianas, se ha planteado la construcción de varios modelos PLS-DA. La finalidad es encontrar la combinación de variables explicativas que permitan realizar predicciones con la mayor exactitud (balanceada) posible.

Se excluyen las siguientes variables del análisis:

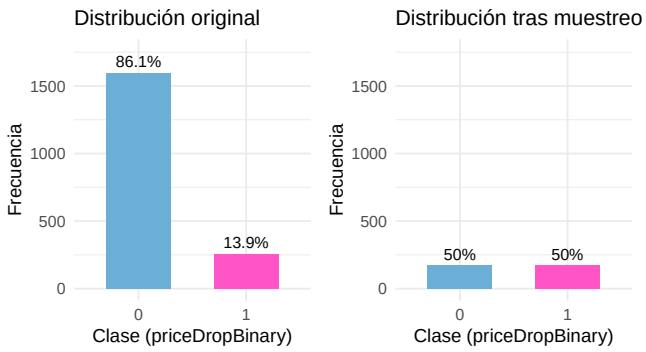
```
-URL, lat, lng, zipCode, creationDate (no aportan información relevante)

-municipality, neighborhood (incluiremos el cluster servicios en su lugar)

-princeAmountDrop (muy relacionada con priceBinaryDrop)
```

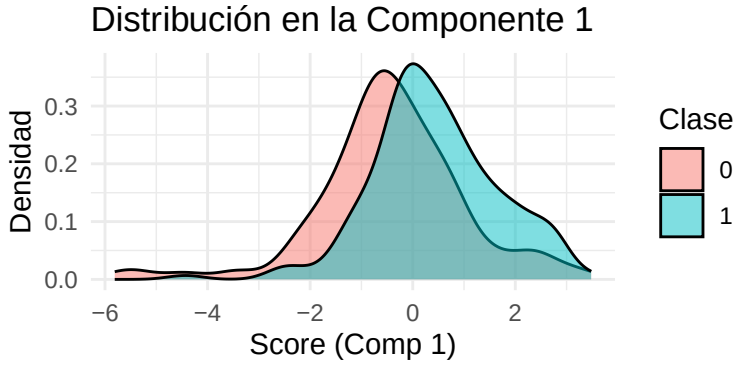
Variables explicativas consideradas en el análisis: [1] “ownerType” “energyEfficiencyRatingType” [3] “environmentImpactRatingType” “bathrooms”
[5] “floor” “hotWater”
[7] “rooms” “surface”
[9] “tieneAscensor” “tieneTrastero”
[11] “tieneCalefaccion” “tieneAireAcondicionado”
[13] “propertySubtypeId” “GeoGeneralRating”
[15] “propertyCounter_buy” “priceDescription_buy”
[17] “propertyCounter_rent” “priceDescription_rent”
[19] “supermarket_count” “pharmacy_count”
[21] “hospital_count” “university_count”
[23] “college_count” “public_transport_count”
[25] “priceAmount” “cluster”

En el gráfico de frecuencias inicial, vemos un claro **desbalanceo de las clases** (86% N y 14% P). Para evitar obtener un modelo sesgado, inclinado a predecir la clase dominante, efectuaremos un *undersampling* y entrenaremos el modelo con una base reducida de **frecuencias uniformes**. Para validar el modelo de forma realista, separaremos entre datos de test y train y, posteriormente, eliminaremos el desbalanceo de la matriz de entrenamiento. En este paso, también escalamos las variables numéricas, puesto que las diferencias en las magnitudes podrían sesgar los resultados del análisis. Una vez separados los datos de prueba y entrenamiento, eliminamos el desbalanceo de estos últimos.



Utilizamos el conjunto de train para generar la primera versión del modelo predictivo. Seguidamente, implementamos la validación cruzada *k-folds* con k=5 y 10 repeticiones para obtener el número de componentes con el que se minimiza el error de predicción. En nuestro modelo pls-DA, el error es menor con 1 única componente principal. No obstante, destacamos el solapamiento entre los inmuebles que bajan de precio y los que no. Esto nos indica que, seguramente, los grupos no están lo suficientemente diferenciados para realizar predicciones acertadas.

```
## Componentes óptimos: 1
```

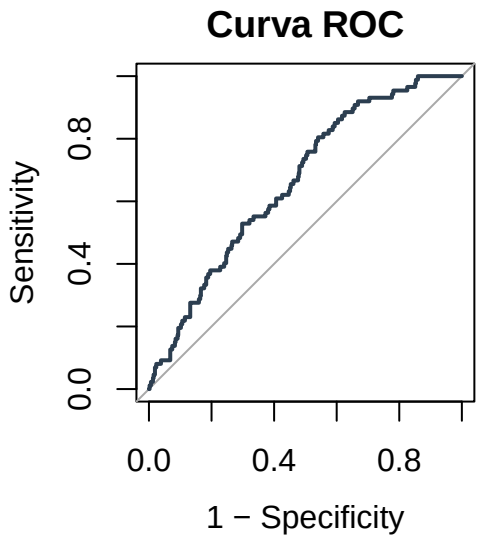


Dado que nuestro objetivo es predecir correctamente si un piso bajará de precio, pero partimos de un conjunto desbalanceado donde la mayoría se mantienen estables, utilizaremos como medidad de calidad la Exactitud Balanceada *(balanced accuracy)*. Esta métrica tiene en cuenta los aciertos en pisos que bajan como en los que no, evitando que el modelo favorezca sistemáticamente la clase mayoritaria. Así, buscamos un rendimiento más justo y representativo que mejore la utilidad práctica del modelo en ambos casos.

```
##          Real
## Predicho   0   1
##        0 300  39
##        1 170  48
```

```
## Accuracy: 0.625
```

```
## Balanced Accuracy: 0.595
```



La curva ROC obtenida para el modelo PLS-DA muestra una capacidad predictiva aceptable, situándose por encima de la línea de referencia (modelo aleatorio). Esto indica que el modelo inicial logra determinar qué pisos bajarán de precio mejor que el azar. Para mejorarlo, calculemos el **umbral de clasificación** que maximiza la **exatitud balanceada**:

```
## Umbral óptimo para máxima Balanced Accuracy: 0.446
```

```
##          Real
## Predicho   0   1
##        0 216  17
##        1 254  70
```

```
## Accuracy: 0.513
```

```
## Balanced Accuracy: 0.632
```

Con el objetivo de **potenciar la capacidad predictiva** efectuaremos un **test de independencia** (t para las variables cuantitativas y chi cuadrada o fisher para las cualitativas); para descartar las variables que no presenten diferencias significativas entre ambas clases. Así, descartaremos la inclusión de variables sin capacidad predictiva, que únicamente generan ruido y disminuyen la fiabilidad de las predicciones.

```
## Variables eliminadas (no presentan diferencias significativas):
```

```
## energyEfficiencyRatingType, environmentImpactRatingType, bathrooms, surface,
## propertyCounter_rent, hospital_count, priceAmount, tieneAscensor,
## tieneTrastero, tieneAireAcondicionado
```

A partir de este punto, el análisis del modelo PLS-DA formado por las variables definitivas se expone con detalle en la **memoria principal**.