Actividad 3: Modelización predictiva

Noviembre_2020

Índice

1. Datos y Estadística descriptiva **2**
- 1.1. Lectura de datos . . . 2
- 1.2. Descriptiva y visualización . . . 2

2. Modelo de regresión lineal **3**
- 2.1. Modelo de regresión lineal simple . . . 3
  - 2.1.1. Calcular . . . 3
  - 2.1.2. Describe las diferencias entre ambos modelos y compáralos. . . . 3
  - 2.1.3. Para cada modelo, realiza un gráfico de dispersión XY e interpretar brevemente el gráfico resultante. . . . 3
- 2.2. Modelo de regresión lineal múltiple (regresores cuantitativos) . . . 3
  - 2.2.1. Calcular . . . 3
  - 2.2.2. Indicar el efecto de cada variable regresora e interpretar el modelo. . . . 3
  - 2.2.3. Evaluar la bondad de ajuste a través del coeficiente de determinación ajustado. . . . 3
  - 2.2.4. Ampliar el modelo anterior con las variables `room_num`, `n_hos_beds` y `n_hot_rooms`. . 3
- 2.3. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos) . . . 3
  - 2.3.1. Aplicar un modelo de regresión lineal múltiple y explicar el resultado. . . . 3
  - 2.3.2. ¿Es significativamente mejor el nuevo modelo? . . . 3
  - 2.3.3. Efectuar una predicción del precio de la vivienda. . . . 3
  - 2.3.4. Efectuar una verificación visual de las suposiciones de modelización. . . . 3

3. Modelo de regresión logística **4**
- 3.1. Regresores cuantitativos . . . 4
  - 3.1.1. Calcular . . . 4
  - 3.1.2. Interpretar . . . 4
- 3.2. Regresores cualitativos . . . 4
  - 3.2.1. Calcular . . . 4
  - 3.2.2. Interpretar . . . 4
- 3.3. Regresores cuantitativos y cualitativos . . . 4
  - 3.3.1. Interpretar . . . 4
  - 3.3.2. Predicción de venta . . . 4
  - 3.3.3. Estimación por resustitución de la precisión del modelo . . . 4
  - 3.3.4. Visualización . . . 5

4. Conclusión **5**

En esta actividad nos planteamos dos objetivos, uno consiste en predecir el precio de venta de una vivienda, y el otro en predecir la expectativa que una vivienda sea vendida. En ambos casos nos interesa conocer los factores influyentes.

Usaremos el fichero de datos `house.csv`. Cuyas variables (se mantiene el nombre original en inglés) son:

Variable	Descripción
price	Precio de venta de una propiedad por parte del propietario
resid_area	Proporción de área residencial en la ciudad
air_qual	Calidad del aire de ese vecindario
room_num	Número medio de habitaciones en casas de esa localidad

Variable	Descripción
age	Años de la construcción inmobiliaria
dist1	Distancia del centro de empleo 1
dist2	Distancia del centro de empleo 2
dist3	Distancia del centro de empleo 3
dist4	Distancia del centro de empleo 4
teachers	Número de maestros por cada mil habitantes en el municipio
poor_prop	Proporción de población pobre en la ciudad
airport	Hay un aeropuerto en la ciudad
n_hos_beds	Número de camas de hospital por mil habitantes en la ciudad
n_hot_rooms	Número de habitaciones de hotel por cada mil habitantes de la ciudad
waterbody	Qué tipo de fuente natural de agua dulce hay en la ciudad
rainfall	Precipitación media anual en centímetros
bus_ter	Hay una terminal de autobuses en la ciudad
parks	Proporción de terrenos asignados como parques y áreas verdes en la ciudad
Sold	Si la propiedad se vendió (1) o no (0)

Aspectos importantes a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- El nivel de confianza por defecto es del 95 %, a no ser que se especifique un valor diferente.
- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.

1. Datos y Estadística descriptiva

1.1. Lectura de datos

En primer lugar, leed el fichero de datos y verificad que los tipos de datos se interpretan correctamente. Si fuera necesario, haced las oportunas conversiones de tipos.

1.2. Descriptiva y visualización

A continuación, comenzaremos el estudio descriptivo, para caracterizar el tipo de variables, detectar posible datos faltantes, outliers, variables con varianza nula o casi nula, etc.

2. Modelo de regresión lineal

2.1. Modelo de regresión lineal simple

2.1.1. Calcular

Estimar por mínimos cuadrados ordinarios dos modelos lineales que expliquen la variable `price`, uno en función de la variable `teachers` y otro en función de la variable `poor_prop`.

2.1.2. Describe las diferencias entre ambos modelos y compáralos.

2.1.3. Para cada modelo, realiza un gráfico de dispersión XY e interpretar brevemente el gráfico resultante.

2.2. Modelo de regresión lineal múltiple (regresores cuantitativos)

2.2.1. Calcular

Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable `price` en función de `age`, `teachers`, `poor_prop`.

2.2.2. Indicar el efecto de cada variable regresora e interpretar el modelo.

2.2.3. Evaluar la bondad de ajuste a través del coeficiente de determinación ajustado.

2.2.4. Ampliar el modelo anterior con las variables `room_num`, `n_hos_beds` y `n_hot_rooms`.

Comparar los dos modelos. ¿Es significativamente mejor el nuevo modelo?

2.3. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)

Queremos conocer en qué medida el modelo anterior (Modelo 2.2) se ve afectado por la inclusión de la variable `airport`.

2.3.1. Aplicar un modelo de regresión lineal múltiple y explicar el resultado.

2.3.2. ¿Es significativamente mejor el nuevo modelo?

2.3.3. Efectuar una predicción del precio de la vivienda.

Para una vivienda cuyas características son:

`age` =70, `teachers` =15 , `poor_prop` =15, `room_num` =8, `n_hos_beds`=8, `n_hot_rooms`=100

Utilizar el modelo Model.2.2

2.3.4. Efectuar una verificación visual de las suposiciones de modelización.

Analiza los residuos del modelo. Comenta los resultados.

3. Modelo de regresión logística

Se desea ajustar un modelo predictivo para predecir la expectativa que una vivienda sea vendida y conocer los factores influyentes en la predicción.

Convertir la variable `Sold` a tipo *factor* y recodificar los valores, asignando "Not" al 0 y "Yes" al 1.

3.1. Regresores cuantitativos

3.1.1. Calcular

Estimar el modelo de regresión logística donde la variable dependiente es `Sold` y las explicativas `price`, `age`, `poor_prop`

3.1.2. Interpretar

Estima los odds ratio de las variables `price`, `age`, `poor_prop` mediante un intervalo de confianza del 95 % e interpreta los intervalos obtenidos. ¿Cuál sería el odds ratio de un quinquenio?

3.2. Regresores cualitativos

3.2.1. Calcular

Estimar el modelo de regresión logística donde la variable dependiente es `Sold` y la explicativa `airport`

3.2.2. Interpretar

Estima el odds ratio de la variable `airport` mediante un intervalo de confianza del 95 % e interpreta el intervalo obtenido.

3.3. Regresores cuantitativos y cualitativos

Estimar el modelo de regresión logística donde la variable dependiente es `Sold` y los regresores `price`, `age`, `poor_prop` y `airport`.

3.3.1. Interpretar

Estima los odds ratio de las variables regresoras mediante un intervalo de confianza del 95 % e interpreta los intervalos obtenidos. ¿Qué regresor tiene más impacto en la probabilidad de venta?

3.3.2. Predicción de venta

Para una vivienda cuyas características son:

`price`=20, `age=50`, `poor_prop=50` y `airport=` YES.

3.3.3. Estimación por resustitución de la precisión del modelo

Proporcionar la tabla de confusión correspondiente al modelo. Comenta los resultados.

3.3.4. Visualización

Para los distintos valores de la variable `price = c(20,30,40)` se representaran las tres series de probabilidades de venta en un mismo gráfico de dispersión XY. En concreto, para cada valor de `price`, se tomarán los valores fijos de `age` =50, `airport =` "YES", y se representarán las probabilidades de venta (eje Y) para los valores de `poor_prop = c(5,25,35,50,65)` (eje X). Comenta el gráfico obtenido.

4. Conclusión

Recopilar las conclusiones alcanzadas en los apartados 2.1, 2.2, 2.3, 3.1, 3.2 y 3.3. En cada caso, puedes acompañar tus conclusiones con los niveles de confianza y/o los p-valores correspondientes.

Puntuación

- Apartado 1 (10 %)
- Apartado 2.1 (10 %)
- Apartado 2.2 (10 %)
- Apartado 2.3 (20 %)
- Apartado 3.1 (10 %)
- Apartado 3.2 (10 %)
- Apartado 3.3 (20 %)
- Apartado 4 (5 %)
- Calidad del informe dinàmico (5 %)