Modelos predictivos

Enunciado

Semestre 2020.2

Índice

Introducción

En esta actividad usaremos un conjunto de datos sobre el aeropuerto internacional de San Francisco (dat_SFO). Ha sido galardonado dos veces, como el mejor aeropuerto en América del Norte. En este estudio se analizarán los datos de vuelos recogidos durante el año 2015. El archivo contiene aproximadamente 145000 registros y 28 variables.

Las principales variables son:

- Month: Día del mes de salida del vuelo.
- Day of week: Día de la semana de salida del vuelo.
- Airline: Nombre en siglas de la compañía aérea.
- Destination Airport: Aeropuerto de destino.
- Scheduled Departure: Hora de salida del vuelo estimada por la compañía.
- Departure Time: Hora de salida real del vuelo.
- Departure Delay: Diferencia entre la hora de salida estimada y la real.
- Air Time: Tiempo real de vuelo en aire,
- Distance: Distancia entre los aeropuertos origen y llegada.
- Scheduled Arrival: Hora de llegada del vuelo estimada por la compañía.
- Arrival Time: Hora de llegada del vuelo

- Arrival Delay: Diferencia entre la hora de llegada estimada y la real.
- Late Aircraft Delay: Retraso por llegada tarde del avión.
- Diverted: Indicador de vuelo desviado, siendo cero si el vuelo se ha efectuado con normalidad y uno si ha sido desviado.
- Cancelled: Indicador de vuelo cancelado, siendo cero si el vuelo se ha efectuado y uno si no.

Cada año una cantidad considerable de vuelos de diferentes aerolíneas se retrasa o cancela, costando al sistema de transporte aéreo miles de millones de euros en pérdidas de tiempo y dinero. En esta actividad se pretende realizar un estudio de los retrasos de los vuelos, tanto en salidas como llegadas.Para ello, se estudiarán las relaciones entre los mismos y varias variables. Primero se estudiarán las relaciones lineales y posteriormente se evaluarán los posibles factores de riesgo de estos retrasos.

A continuación, se especifican los pasos a seguir.En la entrega, se debe respetar la misma numeración de los apartados del índice.

1. Modelo de regresión lineal

1.1. Modelo de regresión lineal (regresores cuantitativos)

a) Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable DEPARTURE_DELAY en función de la variable ARRIVAL_DELAY. Se evaluará la bondad del ajuste, a partir del coeficiente de determinación. Calcular el coeficiente de correlación y explicar su relación con el coeficiente de determinación.

NOTA: En la base de datos los nombres de las variables están en mayúsculas.

b) Se añadirá al modelo anterior la variable independiente DISTANCIA. ¿Existe una mejora del ajuste?. Razonar.

c) Posteriormente, se procederá a dividir la muestra en dos, según los vuelos sean o no más largos. Se tomará por larga distancia aquéllos con un recorrido superior a 600 millas. Razonar los resultados.

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

En este apartado se estudiará la relación de DEPARTURE_DELAY, con las variables explicativas ARRIVAL_DELAY y LATE_AIRCRAFT_DELAY. Para ello se procederá a la recodificación de la variable LATE_AIRCRAFT_DELAY, en en mayor y menor o igual a 15 minutos.

1.3. Diagnosis del modelo

Para la diagnosis se escoge el modelo construído en el apartado b) y se pintarán dos gráficos: uno con los valores ajustados frente a los residuos (que nos permitirá ver si la varianza es constante) y el gráfico cuantil-cuantil que compara los residuos del modelo con los valores de una variable que se distribuye normalmente(QQ plot). Interpretar los resultados.

1.4. Predicción del modelo

Según el modelo del apartado b), calcular el retraso en la salida de un avión, que después de recorrer 2500 millas ha llegado a su destino con 30 minutos más tarde.

2. Modelo de regresión logística.

2.1. Estudio de relaciones entre variables.

Se quiere estudiar la probabilidad que tiene un avión de sufrir un retraso.

Para ello, primero se creará una nueva variable dicotómica llamada **delay_SFO**. Esta nueva variable está relacionada con los valores de la variable Departure_Delay. Se codificará de la siguiente: Si el valor de dicha variable es menor a 15 minutos, se puede asumir que el vuelo no va con retraso y se codificará con el valor 0, en caso contrario, se codificará con el valor 1.

a) Visualizar la relación entre delay_SFO y las variables independientes:DAY_OF_WEEK y AIRLINE. Calcular las frecuencias relativas por fila y columna. Interpretar el significado. Visualizar con barplot.

b) Para comprobar si existe asociación entre las variable dependiente y cada una de las variables explicativas, se aplicará el test Chi-cuadrado de Pearson. Un resultado significativo nos dirá que existe asociación. Interpretar.

2.2. Modelo de regresión logística.

a) Estimar el modelo de regresión logística tomando como variable dependiente delay_SFO y variable explicativa DAY_OF_WEEK. Se tomará como día de referencia el lunes. Se puede considerar que el día de la semana es un factor de riesgo? Justifica tu respuesta.

b) Idem al anterior tomando como variable explicativa AIRLINE. Se tomará como aerolínea de referencia AA. Se puede considerar que la aerolínea es un factor de riesgo? Justifica tu respuesta.

c) Se creará un modelo con la variable dependiente y las variable explicativas DAY_OF_WEEK (la obtenida en el apartado a) y DISTANCE. ¿Se observa una mejora con referencia a los anteriores? Explicar.

d) Se creará un nuevo modelo con la variable dependiente y tomando como variables explicativas, aquéllas que han sido significativas en los apartados anteriores, y además se añadirá la variable ARRIVAL_DELAY.¿Se observa una mejora con referencia a los anteriores? Explicar. Realizad el cálculo de las OR.

2.3. Predicción

Según el modelo del apartado c), calcula la probabilidad de retraso en el vuelo, si nuestro destino está a 1500 millas y viajamos en jueves.

2.4. Bondad del ajuste

Usa el test de Hosman-Lemeshow para ver la bondad de ajuste, tomando el modelo del apartado c). En la librería ResourceSelection hay una función que ajusta el test de Hosmer- Lemeshow.

2.5. Curva ROC

Dibujar la curva ROC, y calcular el área debajo de la curva con los modelos de los apartados c) y d).Discutir el resultado.

3. Conclusiones del análisis

En este apartado se deberán exponer las conclusiones en base a los resultados obtenidos en todo el estudio. Regresión lineal y logística.

Puntuación de los apartados

- Apartado 1.1 (10 %)
- Apartado 1.2 (10 %)
- Apartado 1.3 (10 %)

- Apartado 1.4 (5 %)
- Apartado 2.1 (10 %)
- Apartado 2.2 (15 %)
- Apartado 2.3 (10 %)
- Apartado 2.4 (10 %)
- Apartado 2.5 (10 %)
- Apartado 3 y Calidad del informe dinámico (10 %)