Modelos predictivos

Enunciado

Semestre 2022.1

Índice

Introducción

En esta actividad usaremos un conjunto de datos recogidos en un estudio de satisfacción de clientes de un determinado aeropuerto internacional. En dicho estudio se pidió a los pasajeros que calificaran su nivel de satisfacción en general y de otros aspectos de su vuelo, como comodidad del asiento, facilidad de reserva en línea, limpieza,. También se registraron otras variables como el tiempo de demora en salidas y llegadas, entre otras. El objetivo principal de este estudio es averiguar cuáles son los factores que más influyen en que un pasajero se muestre satisfecho y cuáles no. Este tipo de análisis son muy útiles, ya que facilitan la puesta en marcha de mejoras concretas en los servicios que así lo requieran.

El archivo contiene aproximadamente 129446 registros y 20 variables. Las principales variables son:

- satisfaction: Satisfacción del viajero, medida en dos categorías: (neutral or dissatisfied y satisfied).
- Gender: Sexo biológico.
- Customer_Type: Tipo de cliente.
- Age: Edad.
- Type_Travel: Tipo de viaje.
- Class: Tipo de tarifa.
- Distance: Distancia entre origen y destino.
- Departure_Delay: Tiempo de retraso en la salida, en minutos.
- Arrival_Delay: Tiempo de retraso en la llegada a destino, en minutos.

Y las siguientes variables ordinales medidas en cinco categorías que van desde 1, (Muy poco satisfecho) a 5 (Muy satisfecho).

- Seat_comfort: Comodidad del asiento.
- Food_drink: Comida y bebida.
- Gate: Distancia a la puerta de embarque.
- Wifi: Servicio de Wifi.
- Ent: Entretenimiento.
- Ease_booking: Fácil booking.
- Service: Servicio en general.
- Baggage_handling: Equipaje de mano.
- Checkin_service: Checking.
- Cleanliness: Limpieza.
- Online_boarding: Embarque online.

Primero se estudiarán las posibles relaciones lineales entre el retraso en la llegada de los vuelos a destino `Arrival_Delay` y diferentes variables independientes. En la segunda parte de la actividad se buscarán los posibles factores de riesgo o protección que afectan a la satisfacción de los pasajeros mediante regresión logística.

Según los datos de la encuesta se sabe que el 54,7% de los pasajeros estaban satisfechos con su vuelo. Nos interesa averiguar que es que hace que estén más o menos satisfechos, y así poder mejorar el servicio. A continuación, se especifican los pasos a seguir.

Nota::

- 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.
- Se valorará con un 20% el informe ejecutivo (tabla y resúmen con las principales conclusiones).

1 Regresión lineal

1.1 Modelo de regresión lineal (variables cuantitativas)

Se quiere estudiar si la distancia entre destinos y el retraso de vuelo de salida, influyen o no en que un vuelo no llegue a la hora prefijada. Se pide:

a) Estimad por mínimos cuadrados ordinarios un modelo lineal que explique la variable `Arrival_Delay` en función de la variable `Distance`. Interpretad.

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

1.2 Modelo de regresión lineal (variables cuantitativas y cualitativas).

En este apartado se estudiará si la puntuación dada al servicio y a la comida y bebida, podría variar si un vuelo llega con más o menos retraso. También interesa observar si cuánto mayor sea el retraso puede afectar al grado de satisfacción en general. Por otro lado también se estudiará si ser un cliente leal, pueda o no influir en los retrasos.

a) Se añadirá al modelo del apartado 1.b), las variables cualitativas ordinales `Service`, `Food_drink` y `satisfaction`, junto con la variable cualitatitva nominal `Customer_Type`. A la vista de los resultados, estudiar si son o no significativas. Decidid cuáles de las variables explicativas propuestas hasta el momento deben quedarse en el modelo de regresión lineal. Se le llamará modelo final **ModelF**.

b) Comprobad si existen o no problemas de colinealidad en dicho modelo final **ModelF**.

Nota: La variable `satisfaction` inicialmente tenía 5 valores desde 1, (Muy poco satisfecho) a 5 (Muy satisfecho). Para este estudio se ha agrupado en dos (neutrales o insatisfechos y satisfechos).

1.3 Diagnosis del modelo.

Para la diagnosis se escoge el **ModelF** construído y se piden 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). Interpretad los resultados.

1.4 Predicción del modelo

Según **ModelF**, calculad el retraso en la llegada del vuelo, si un viajero satisfecho ha recorrido una distancia de 2500 millas y ha tenido un retraso en la salida de 30 minutos. Se conoce que dicho viajero a puntuado con 3, su nivel de satisfacción sobre el servicio (`Service`).

2 Regresión logística.

Se quiere estudiar cuáles son los factores que más influyen en el grado de satisafacción de los pasajeros de avión.

Para ello, primero se creará una nueva variable dicotómica llamada **satisfaction_re**. Esta nueva variable está relacionada con los valores de la variable `satisfaction`. Se codificará de la siguiente forma: "neutral or dissatisfied" toma el valor 0 y "satisfied" el valor 1. (esta variable puede crearse desde el inicio de la actividad).

Se pide estimar un modelo de regresión logística tomando como variable dependiente `satisfaction_re` y un conjunto de variable explicativas de la base de datos, que se detallarán posteriormente.

Para poder estimar de forma más objetiva la precisión del modelo, separaremos el conjunto de datos en dos partes: el conjunto de entrenamiento (training) y el conjunto de prueba (testing). Ajustaremos el modelo de regresión logística con el conjunto de entrenamiento, y evaluaremos la precisión con el conjunto de prueba.

Se pide:

2.1 Generación de los conjuntos de entrenamiento y de test

Generad los conjuntos de datos para entrenar el modelo (training) y para testarlo (testing). Se puede fijar el tamaño de la muestra de entrenamiento a un 80% del original.

2.2 Estimación del modelo con el conjunto de entrenamiento e interpretación

Tomando como base, training:

a) Estimad el modelo de regresión logística siendo la variable dependiente `satisfaction_re` y tomando todas las variables explicativas de la base de datos. Tened en cuenta la variable `satisfaction` sin recodificar no puede ser una variable explicativa.

b) Estudiad la presencia o no de colinealidad. En el caso de existir, eliminar la variable o variables que consideréis.

c) Una vez corregido el modelo por la presencia o no de colinealidad, se pide:

- Interpretad la salida del modelo final. Se le llamará **ModlgF**.
- Resumid cuáles de las variables pueden considerarse factores de riesgo o protección.

2.3 Cálculo de las OR (Odds-Ratio)

Según los resultados de **ModlgF**, calculad las OR, correspondientes. Interpretad los valores de las OR de las variables explicativas siguientes: `Class`,`Customer_Type`, `Gender` y `Ent`.

2.4 Matriz de confusión

A continuación analizad la precisión de **ModlgF**, comparando la predicción del modelo contra el conjunto de prueba (testing).Se asumirá que la predicción del modelo es 1, *satisfied*, si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 en caso contrario. Analizad la matriz de confusión y las medidas de sensibilidad y especificidad.

2.5 Predicción

Según **ModlgF**, calculad la probabilidad de que el cliente encuestado número tres (tercera fila de la base de datos) estuviera o no satisfecho con la aerolínea.

2.6 Bondad del ajuste

a) Evaluad la bondad del ajuste, mediante la *devianza*. Para que **ModlgF** sea bueno la devianza residual debe ser menor que la devianza nula. En ese caso el modelo predice la variable dependiente con mayor precisión.

b) Evaluad la eficacia del modelo según el test Chi-cuadrado. En este caso el valor del estadístico Chi-cuadrado observado es igual a la diferencia de devianzas (nula-residual). Calculad la probabilidad asociada al estadístico del contraste utilizando la función **pchisq**.

Nota: los grados de libertad se calcularán como la diferencia de grados de libertad entre el modelo nulo y residual.

2.7 Curva ROC

Dibujad la curva ROC y calcular el área debajo de la curva con **Modlg**. Discutid el resultado.

3 Informe Ejecutivo

3.1 Presentación de los principales resultados del estudio en una tabla

Presentad una tabla, de tal forma que en cada fila se detallen los resultados principales de cada apartado.

Ejemplo correspondiente al apartado 1: la pregunta de investigación planteada, los resultados obtenidos (la significación o no de las variables explicativas, valores del coeficiente de determinación„) y la conclusiones. Se podría tomar un formato similar a la actividad anterior.

3.2 Resúmen ejecutivo. Conclusiones del análisis

Resumid las conclusiones del estudio para una audiencia no técnica, indicando las respuestas a las preguntas de investigación planteadas. El resumen no debe ocupar más de media página.

Nota: esta pregunta trabaja la competencia de comunicación que es muy importante en el rol de analista de datos.

Puntuación de los apartados

- Apartado 1.1 (5%)
- Apartado 1.2 (10%)
- Apartado 1.3 (5%)
- Apartado 1.4 (5%)
- Apartado 2.1 (10%)
- Apartado 2.2 (10%)
- Apartado 2.3 (10%)
- Apartado 2.4 (5%)
- Apartado 2.5 (5%)
- Apartado 2.6 (5%)
- Apartado 3 (20%)
- calidad del informe dinámico (10%)