Tarea de regresión

1. Predicción de tarifas aéreas en nuevas rutas.

El siguiente problema se produce en Estados Unidos a finales de los años 90, cuando muchas de las principales ciudades del país se enfrentaban a problemas de congestión aeroportuaria, en parte como resultado de la desregulación de las aerolíneas en 1978. Tanto las tarifas como las rutas se liberaron de la regulación, y las aerolíneas de bajo costo como Southwest (SW) empezaron a competir en las rutas existentes y a iniciar servicios sin escalas en rutas que anteriormente carecían de ellos.

La construcción de aeropuertos completamente nuevos no suele ser viable, pero a veces las bases militares desmanteladas o los aeropuertos municipales más pequeños se pueden reconfigurar como aeropuertos regionales o comerciales más grandes. Hay numerosos actores e intereses involucrados en el problema (líneas aéreas, autoridades municipales, estatales y federales, grupos cívicos, el ejército, operadores aeroportuarios), y una empresa de consultoría de aviación está buscando contratos de asesoramiento con estos actores.
La empresa necesita modelos predictivos para respaldar su servicio de consultoría. Una cosa que la empresa podría querer es predecir las tarifas, en caso de que se ponga en servicio un nuevo aeropuerto. La empresa comienza con el archivo Airfares.csv, que contiene datos reales que se recopilaron entre el tercer trimestre de 1996 y el segundo trimestre de 1997.

Las variables de estos datos se enumeran en la tabla 1 y se cree que son importantes para predecir el precio de las tarifas. Hay algunos datos de aeropuerto a aeropuerto disponibles, pero la mayoría de los datos se encuentran a nivel de ciudad a ciudad. Una pregunta que será de interés en el análisis es el efecto que tiene la presencia o ausencia de Southwest en el precio de las tarifas.

S_CODE	Starting airport's code
S_CITY	Starting city
E_CODE	Ending airport's code
E_CITY	Ending city
COUPON	Average number of coupons (a one-coupon flight is a nonstop flight, a two-coupon flight is a one-stop flight, etc.) for that route
NEW	Number of new carriers entering that route between Q3-96 and Q2-97
VACATION	Whether (Yes) or not (No) a vacation route
SW	Whether (Yes) or not (No) Southwest Airlines serves that route
HI	Herfindahl index: measure of market concentration
S_INCOME	Starting city's average personal income
E_INCOME	Ending city's average personal income
S_POP	Starting city's population
E_POP	Ending city's population
SLOT	Whether or not either endpoint airport is slot-controlled (this is a measure of airport congestion)
GATE	Whether or not either endpoint airport has gate constraints (this is another measure of airport congestion)
DISTANCE	Distance between two endpoint airports in miles
PAX	Number of passengers on that route during period of data collection
FARE	Average fare on that route

a. Explora los predictores numéricos y la respuesta (FARE) creando una tabla de correlación y examinando algunos diagramas de dispersión entre FARE y esos predictores. ¿Cuál parece ser el mejor predictor individual de FARE?

b. Explora los predictores categóricos (excluyendo los primeros cuatro) calculando el porcentaje de vuelos en cada categoría. Crea una tabla dinámica con la tarifa promedio en cada categoría. ¿Qué predictor categórico parece mejor para predecir el precio de las tarifas?

c. Encuentra un modelo para predecir la tarifa promedio en una nueva ruta:

i. Convierte las variables categóricas (por ejemplo, SW) en variables ficticias. Luego, divide los datos en conjuntos de entrenamiento y validación. El modelo se ajustará a los datos de entrenamiento y se evaluará en el conjunto de validación.
ii. Utiliza la regresión por pasos para reducir la cantidad de predictores. Puedes ignorar los primeros cuatro predictores (S_CODE, S_CITY, E_CODE, E_CITY). Informa el modelo estimado seleccionado.
iii. Repite (ii) utilizando una búsqueda exhaustiva en lugar de una regresión por pasos. Compara el mejor modelo resultante con el que obtuviste en (ii) en términos de los predictores que están en el modelo.
iv. Compara la precisión predictiva de ambos modelos (ii) y (iii) utilizando medidas como RMSE y gráficos de error promedio y elevación.
v. Utilizando el modelo (iii), predecir la tarifa promedio en una ruta con las siguientes características: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28,760, E_INCOME = $27,664, S_POP = 4,557,004, E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 millas.
vi. Predecir la reducción en la tarifa promedio en la ruta en (v) si Southwest decide cubrir esta ruta [utilizando el modelo (iii)].
vii. En realidad, ¿cuál de los factores no estará disponible para predecir la tarifa promedio desde un nuevo aeropuerto (es decir, antes de que comiencen a operar vuelos en esas rutas)? ¿Cuáles se pueden estimar? ¿Cómo?
viii. Selecciona un modelo que incluya únicamente los factores que están disponibles antes de que comiencen a operar los vuelos en la nueva ruta. Utilice una búsqueda exhaustiva para encontrar dicho modelo.
ix. Utiliza el modelo en (viii) para predecir la tarifa promedio en una ruta con características COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28,760, E_INCOME = $27,664, S_ POP = 4,557,004, E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12782, DISTANCE = 1976 millas.

x. Compara la precisión predictiva de este modelo con el modelo (iii). ¿Es este modelo lo suficientemente bueno o vale la pena reevaluarlo una vez que comiencen los vuelos en la nueva ruta?

d. En industrias competitivas, un nuevo participante con un plan de negocios novedoso puede tener un efecto disruptivo en las empresas existentes. Si el modelo de negocios de un nuevo participante es sostenible, otros participantes se ven obligados a responder modificando sus prácticas comerciales. Si el objetivo del análisis fuera evaluar el efecto de la presencia de Southwest Airlines en la industria de las aerolíneas en lugar de predecir las tarifas en nuevas rutas, ¿en qué se diferenciaría el análisis? Describe los aspectos técnicos y conceptuales.

2. Situación financiera de los bancos.

El archivo Banks.csv incluye datos sobre una muestra de 20 bancos. La columna "Condición financiera" (Financial Condition) registra el juicio de un experto sobre la situación financiera de cada banco. Esta variable de resultado toma uno de dos valores posibles: débil o fuerte según la situación financiera del banco. Los predictores son dos ratios utilizados en el análisis financiero de los bancos: TotLns&Lses/Assets es la relación entre los préstamos y arrendamientos totales y los activos totales y TotExp/Assets es la relación entre los gastos totales y los activos totales. El objetivo es utilizar los dos ratios para clasificar la situación financiera de un nuevo banco.

Ejecuta un modelo de regresión logística (sobre todo el conjunto de datos) que modele el estado de un banco como una función de las dos medidas financieras proporcionadas. Especifica la clase de éxito como débil (esto es similar a crear una variable ficticia que sea 1 para los bancos financieramente débiles y 0 en caso contrario) y utiliza el valor de corte predeterminado de 0.5.

a. Escribe la ecuación estimada que asocia la situación financiera de un banco con sus dos predictores en tres formatos:

i. El logit como función de los predictores
ii. Las probabilidades (odds) como función de los predictores
iii. La probabilidad como función de los predictores

b. Considera un banco nuevo cuya relación préstamos totales y arrendamientos/activos = 0.6 y relación gastos totales/activos = 0.11. A partir de tu modelo de regresión logística, estima las siguientes cuatro cantidades para este banco (usa R para hacer todos los cálculos intermedios; muestra tus respuestas finales con cuatro decimales): el logit, las probabilidades

(odds), la probabilidad de ser financieramente débil y la clasificación del banco (use el valor de corte = 0.5).

c. El valor de corte de 0.5 se utiliza junto con la probabilidad de ser financieramente débil. Calcula el umbral que se debe utilizar si queremos hacer una clasificación basada en las probabilidades de ser financieramente débil y el umbral para el logit correspondiente.

d. Interpreta el coeficiente estimado para la relación entre préstamos y arrendamientos totales y activos totales (TotLns&Lses/Assets) en términos de las probabilidades de ser financieramente débil.

e. Cuando un banco que está en malas condiciones financieras se clasifica erróneamente como financieramente fuerte, el costo de la clasificación errónea es mucho mayor que cuando un banco financieramente fuerte se clasifica erróneamente como débil. Para minimizar el costo esperado de la clasificación errónea, ¿debería aumentarse o reducirse el valor de corte para la clasificación (que actualmente es de 0.5)?