Evaluación - Regresión Logística

Diplomado en Data Science UC

El archivo *rrhh.csv* contiene información de la situación laboral de los trabajadores de la empresa ABAC. Las variables disponibles son:

- **Estado**: Estado actual del trabajador en la empresa (1: desvinculado, 0: vinculado).
- **Edad**: Edad del trabajador en años.
- **Ratio.Pago**: Medida de pago por hora (numerico)
- **Salario**: Salario mensual en dólares que tiene o tenía el trabajador
- **Dias.trabajados**: Días que lleva o llevaba trabajando en la empresa
- **Ausencias**: Días que ha faltado a trabajar
- **Sexo**: Sexo del trabajador (Female , Male)
- **Estado.Civil**: Estado civil del trabajador (1: divorciado, 2: casado,3: separado, 4: soltero, 5: viuda)
- **Departamento**: Lugar de trabajo en la empresa (Admin Offices,..)
- **Posicion**: Cargo del trabajador/empleado (Accountant I ,....)
- **Desempeño**: Clasificación del desempeño del trabajador.

Carga, limpieza y formato de los datos

a. Cargue los datos en R y revise los formatos de cada variable, recuerde codificar las variables como numéricas o factores según corresponda.

Análisis descriptivo y exploratorio de datos

b. Realice un análisis descriptivo de sus datos. Determinar si existen observaciones faltantes, en el caso de existir tome la decisión de omitirlas del estudio u omitir la variable. Evalúe si existen posibles incongruencias en la fuente de datos (ej: edades negativas). Y finalmente análice la presencia de valores atípicos en las variables. Comente.

c. Realice análisis de cómo se relacionan las variables **continuas** con la variable de interés. Acompañe con gráficos y estadísticas. ¿Qué variables pudieran resultar significativas a la hora de modelar la probabilidad de que el trabajador sea desvinculado a la empresa?

d. Realice análisis de cómo se relacionan las variables **categóricas** con la variable de interés. Acompañe con gráficos y estadísticas. ¿Qué variables pudieran resultar significativas a la hora de modelar la probabilidad de que el trabajador sea desvinculado a la empresa?

Modelamiento

e. Realice una separación de la base de datos en un set de entrenamiento y set de validación, utilice una proporción de 75:25 respectivamente. Para poder replicar sus resultados, fije una semilla antes de obtener los indices. Para ello, utilice la función set.seed(2021).

f. Con los datos de entrenamiento ajuste un modelo de regresión logística para estudiar la probabilidad de que el trabajador sea desvinculado de la empresa. Para ello, utilice las variables edad y desempeño.

g. Calcule e interprete los OR correspondientes al modelo, ¿son estos factores protectores o agravantes de la desvinculación del trabajador?

h. Utilizando un método automatizado, encuentre el modelo óptimo usando como criterio el criterio de información de Akaike (AIC). La función step() puede ser de utilidad.

i. Si usted trabaja en la empresa ABAC, calcule su probabilidad de ser desvinculado. Suponga que sus características son:

- Edad: 34
- Ratio.Pago: 34.95
- Salario: 3345.2
- Dias.trabajados: 3247

- Ausencias: 16
- Sexo: Female
- Estado.Civil: 2
- Departamento: Admin Offices
- Posicion: Sr. Accountant
- Desempeño: Fully Meets

Validación del modelo

j. Utilizando la base de validación y el modelo obtenido en la pregunta anterior, calcule las probabilidades de que el trabajador sea desvinculado.

k. Identifique el punto de corte que optimice la sensibilidad del modelo, pero que cometa como máximo una tasa de falsos positivos (1 - Especificidad) de a lo más un 25%. Use el argumento `returnSensitivityMat = TRUE` en la función `plotROC()`. Y obtenga las matrices de confusión y los indicadores de:

- Sensibilidad
- Especificidad
- Precisión

l. Evalúe el modelo y concluya. Para ello, obtenga e interprete los siguientes estadísticos:

- Área bajo la curva ROC
- Test de Kolmogorov - Smirnov (*Hint: utilice la* `función ks.test(x, y)`).
- Test de Hosmer - Lemeshow (*Hint: utilice la función* `ResourceSelection::hoslem.test()`).