DEPARTAMENTO DE
INGENIERÍA
INFORMÁTICA
UNIVERSIDAD DE SANTIAGO DE CHILE

EJERCICIO PRÁCTICO 15: REGRESIÓN LINEAL Y LOGÍSTICA CON CARET

CONTEXTO

Conocemos varias herramientas que facilitan la búsqueda y construcción de modelos de regresión lineal y logística, además del proceso iterativo para conseguir un modelo confiable.

El objetivo de este ejercicio es utilizar herramientas del paquete caret para automatizar la búsqueda y evaluación de modelos de regresión, puesto que este paquete les será útil en el siguiente curso de la línea de ciencia de datos.

OBJETIVOS DE APRENDIZAJE

1. Preparar una muestra de datos para la construcción de modelos de regresión.
2. Utilizar herramientas del paquete caret para seleccionar predictores, construir y evaluar modelos de regresión.

ÉXITO DE LA ACTIVIDAD

El equipo es capaz de encontrar modelos de regresión confiables, que usan un conjunto reducido de buenos predictores, y de buen desempeño al predecir una variable dependiente.

ACTIVIDADES

Para esta actividad usaremos los datos de medidas anatómicas recolectados por Heinz et al. (2003) que ya conocimos en los ejercicios prácticos anteriores, añadiendo las variables ICM y EN creada en el ejercicio práctico anterior.

1. El equipo copia el enunciado del problema asignado como comentarios de un script R.
2. El equipo lee el enunciado, descarga el archivo de datos (EP13 Datos.csv) desde UVirtual y selecciona las columnas para trabajar de acuerdo con las instrucciones.
3. El equipo construye los modelos solicitados usando la muestra correspondiente.
4. El equipo sube el script con las actividades anteriores comentando en detalle los pasos seguidos.

Fuera del horario de clases, cada equipo debe subir el script realizado UVirtual con el nombre "EP15-respuesta-grupo-i", donde i es el número de grupo asignado. Las respuestas deben subirse antes de las 23:30 del lunes 27 de junio.

PREGUNTA (todos los grupos)

1. Definir la semilla a utilizar, que corresponde a los primeros cinco dígitos del RUN del integrante de mayor edad del equipo.
2. Seleccionar una muestra de 100 personas, asegurando que la mitad tenga estado nutricional “sobrepeso” y la otra mitad “no sobrepeso”.
3. Usando las herramientas del paquete leaps, realizar una búsqueda exhaustiva para seleccionar entre dos y ocho predictores que ayuden a estimar la variable Peso (Weight), obviamente sin considerar las nuevas variables IMC ni EN, y luego utilizar las funciones del paquete caret para construir un modelo de regresión lineal múltiple con los predictores escogidos y evaluarlo usando bootstrapping.

4. Haciendo un poco de investigación sobre el paquete caret, en particular cómo hacer Recursive Feature Elimination (RFE), construir un modelo de regresión lineal múltiple para predecir la variable IMC que incluya entre 10 y 20 predictores, seleccionando el conjunto de variables que maximice R2 y que use cinco repeticiones de validación cruzada de cinco pliegues para evitar el sobreajuste (obviamente no se debe considerar las variables Peso, Estatura ni estado nutricional –Weight, Height, EN respectivamente).
5. Usando RFE, construir un modelo de regresión logística múltiple para la variable EN que incluya el conjunto, de entre dos y seis, predictores que entregue la mejor curva ROC y que utilice validación cruzada dejando uno fuera para evitar el sobreajuste (obviamente no se debe considerar las variables Peso, Estatura –Weight y Height respectivamente– ni IMC).
6. Pronunciarse sobre la confiabilidad y el poder predictivo de los modelos.

CRITERIOS DE EVALUACIÓN

- Obtienen una muestra de datos para poder crear y evaluar modelos de regresión, cumpliendo las restricciones indicadas en el enunciado (semilla, tamaño, balanceo, etc.).
- Escriben código en R correcto que utiliza funciones de los paquetes leaps y caret para explorar el conjunto válido de potenciales predictores, según las instrucciones del enunciado, y que escoge un subconjunto de ellos, respetando el tamaño indicado, de acuerdo al criterio de selección solicitado y la técnica especificada para evitar el sobreajuste.
- Construyen correctamente un modelo de regresión pertinente para predicir la variable de salida indicada, usando como predictores las variables seleccionadas.
- Escriben comentarios y código en R correcto que verifica las condiciones que garantizan que el modelo obtenido tiene un buen nivel de ajuste, es generalizable y tiene buena calidad predictiva, interpretando explícita y correctamente los resultados obtenidos en cada paso y tomando acciones correctivas apropiadas de ser necesarias o comentando los riesgos asociados.
- Entregan conclusiones correctas y completas, basadas en las evaluaciones realizadas y el proceso de búsqueda de predictores seguido, respecto del modelo obtenido.
- El script está completo, ordenado y bien indentado, logrando un programa que es fácil de seguir y que no requiere cambios para que funcione.
- El script está comentado paso a paso, con claridad (basta una lectura para entender) y con buena redacción y ortografía (<= 5 errores), usando vocabulario propio de la disciplina y el contexto del problema.