Ejercicio 1

Elabore una función que permita obtener una valoración completa y personalizada de un modelo GLM. La función, por tanto, deberá obtener:

- Análisis exploratorio, mediante tablas y gráficos, de la relación entre las variables independientes y la dependiente. Utilice para ello plot y gráficos de correlación, diagramas de caja, de densidad,… Nota: distinguir, en función del tipo de la variable dependiente, el tipo de modelo, regresión o clasificación (funciones is.factor o is.numeric).
- Una salida convenientemente formateada de los coeficientes estimados por el modelo, desviaciones estándar, estadístico t y sus p-valores, tanto en su versión clásica (bajo los supuestos del modelo) como Bootstrap.
- Ídem para las medidas de bondad del ajuste. Considere el R^2, el R^2 ajustado y el ECM para la regresión y el error de clasificación y los pseudo R^2 para clasificación.
- Resultados de los tests diagnósticos del modelo: Shapiro-Wilk, Lilliefors, Levene/Brown-Forsythe, White, Breush-Godfrey, Durbin-Watson… determinando si los residuos son normales, homocedásticos y aleatorios. Nota: mediante plot(modelo) ya se obtienen los tests gráficos, pero puede personalizar la presentación usando ggplot2 o cualquiera otra librería gráfica (lattice, plotrix, plotly,...). Dados los resultados obtenidos, la función devolverá un mensaje aconsejando utilizar o no el uso de los estimadores clásicos o los de Bootstrap.
- Incluya un parámetro adicional, indicando si se desea realizar o no una selección de variables y, de realizarse, de qué tipo: forward, backward, stepwise. Utilice como criterio el estadístico BIC.

Nota: para la obtención de los estadísticos Bootstrap puede basarse en los tres paquetes implementados en R: rsample, bootstrap o boot, o realizar una función propia al efecto. Para problemas de clasificación, realice el Bootstrap para cada clase por separado. Puede aprovechar las funciones de estadística descriptiva hechas para el preprocesado en el ejercicio anterior.

Utilice el data frame BostonHousing del paquete mlbench para probar la función en el caso de regresión, y para clasificación la ya preprocesada del ejercicio anterior.

Ejercicio 2

En el documento relativo a Introducción al Machine Learning se elabora un ejemplo de Minería de Datos utilizando la base de datos de german credit balanceada con 750 registros, 375 de cada clase. Uno de los modelos estimados ha sido un modelo Logit, obteniendo un área bajo la curva ROC de 0,771, muy próxima a la del mejor modelo, el Random Forest, con un valor de 0,796.

Es de esperar que un modelo GAM mejore el ajuste del GLM, pero ¿superará al Random Forest? Para ello, elabore un GAM estudiando qué tipo de regresión utilizar (ajuste

polinómico o curvilíneo, Loess, spline de regresión o spline de suavizado) con cada una de las 3 variables independientes cuantitativas: duración, montante y edad. Nota: la implementación de caret no permite actualmente ajustar adecuadamente este tipo de modelos.

Al igual que en el resto de los ejemplos, utilice una estimación Bootstrap con 30 repeticiones para obtener las 5 métricas planteadas en el documento. Compare estadísticamente los resultados. Nota: puede calcular IC de AUC con la librería pROC, funciones auc() y ci.auc(). El resto de métricas puede obtenerse a través de la matriz de confusión obtenida a partir de las réplicas Bootstrap (ojo, en el documento se muestra la matriz de confusión de los 750 datos, no de las réplicas).

Ejercicio 3

La base de datos **environmental** de la librería lattice presenta las mediciones diarias de concentración de ozono, velocidad del viento, temperatura y radiación solar en la ciudad de Nueva York de mayo a septiembre de 1973. En concreto, se dispone de 111 observaciones con la siguiente información:

- Ozono. Concentración promedio de ozono (mediciones por hora) en partes por billón.
- Radiación. Radiación solar (de 08:00 a 12:00) en langleys.
- Temperatura. Temperatura máxima diaria en grados Fahrenheit.
- Viento. Velocidad promedio del viento (a las 07:00 y 10:00) en millas por hora.

En base a otros estudios, los coeficientes beta o, lo que es lo mismo, el aumento de la concentración de ozono según se varíen las condiciones climáticas en una unidad, pueden acotarse del siguiente modo (aproximadamente a un 95% de confianza): radiación 0,05 a 0,15, temperatura 1,3 a 1,7 y viento -3,7 a -3,3. Supondremos la no existencia de covarianzas.

Para la precisión (inversa de la varianza residual), considere como parámetros de la función gamma las dos siguientes: c0 = 56 y d0 = 24649.

Estime un segundo modelo sin incluir estimaciones a priori de los coeficientes beta y sus varianzas, considerando c0 = d0 = 100.

Estime un tercero sin incorporar información a priori. Comente cuál es la parametrización por defecto y razone lo que esto implica.

Compare los resultados con un modelo de regresión lineal clásico o frecuentista. Nota: la esperanza y varianza del error residual en la regresión lineal clásica es: $E(S_R^2) = \sigma^2$ y $Var(S_R^2) = {2\sigma^4}/{n-k-1}$

Por último, estime un modelo Loess, compárelo con el resto y establezca las pertinentes conclusiones. Determine, razonadamente, cuál de ellos sería el de mayor desempeño.