Tarea 3

Ejercicio 2

En nuestro equipo hicimos una búsqueda de diferentes modelos para la predicción del padecimiento de diabetes en los pacientes.
Calificamos diferentes modelos tales como:
-Regresión logit, probit, cauchit, cloglog con todas las interacciones
-Regresión logit, probit, cauchit, cloglog sin interacciones
-Regresión logit, todas las interacciones, selección lasso, lambda.min
-Regresión logit, todas las interacciones, selección lasso, lambda.1se
-Naive Classifier
-LDA y QDA
-KNN, con tunning
-Random Forest, sin tunning

A continuación mostramos una tabla donde se muestran los mejores modelos obtenidos (los demás, aunque los calificamos fueron muy pobres en comparación).
Mostramos su precisión global, su sensibilidad (recall) y su especificidad.

model	accuracy	recall	specificity
Logit, sin interacciones	0.7841026	0.5676923	0.8923077
Random Forest, sin tunning	0.7838462	0.5869231	0.8823077
LDA	0.7830769	0.5707692	0.8892308
Probit, sin interacciones	0.7815385	0.5607692	0.8919231
Todas las interacciones, lasso, lambda.min	0.7771795	0.5315385	0.9000000
QDA	0.7710256	0.5892308	0.8619231
Todas las interacciones, lasso, lambda.1se	0.7700000	0.4653846	0.9223077
Naive Classifier	0.7641026	0.6353846	0.8284615
KNN	0.7602564	0.5007692	0.8900000

Como podemos apreciar, el modelo con mayor precisión global es la regresión logit sin interacciones, sin embargo Random Forest, LDA y probit sin interacciones no están muy lejos en comparación.
Por otro lado, el mejor modelo en cuanto a especificidad, es la regresión logit con lambda.min. Es decir, si quisiéramos reducir la mayor cantidad de falsos negativos podríamos usar este modelo sin ningún problema, ya que si futuras observaciones son clasificadas como "negativo" bajo este modelo, el 90% de las veces habremos clasificado de manera correcta. Es decir, si llega un paciente nuevo y lo ponemos bajo este modelo y resulta ser un negativo, lo más probable es que este nuevo paciente no tenga diabetes. Si sale positivo podemos considerar mejor otro modelo:
Respecto a la mejor sensibilidad tenemos que Naive Classifier es el ganador por relativamente bastante. Es decir, desafortunadamente todos los demás modelos no están muy lejos de "un volado", y en este modelo es en el único que sobrepasamos el 60%.

Como decisión final, ya que lo que más nos interesaría es saber si un nuevo paciente es positivo a diabetes, no recomendaríamos usar sólo un modelo. Un protocolo que recomendamos usar es:
1. Utilizar el modelo con selección lasso lambda.min.

2a. Si sale negativo podemos estar bastante seguros de que no tiene diabetes. Podríamos considerar otro modelo para estar más seguros y no tendríamos que preocuparnos demasiado.
2b. Si sale positivo, probemos con el modelo de Naive Classifier. Si nuevamente sale positivo es bastante probable que el paciente en efecto tenga diabetes, por lo que la empresa debería proseguir como más convenga. Comenzar tratamientos, gastar en realizar estudios de confirmación más certeros, etc.
3b. Si sale negativo después de Naive Classifier, tampoco debemos cantar victoria ya que no es muy preciso. Lo mejor sería seguir contrastando con otros modelos ya que no convendría tomar una decisión al respecto en este caso.

Variables Con Mayor Poder Predictivo

Finalmente, observando los coefficientes de todos los modelos planteados, pudimos observar que las variables que más efecto tienen en el diagnóstico de diabetes son:
- En primer lugar y por mucho, la genética (pedigree)
- Y en segundo lugar, principalmente porque fueron variables de gran peso en el Random Forest exclusivamente: la glucosa, la insulina y la edad.

Entonces es muy importante tener en cuenta estos factores de los pacientes. Genéticamente no hay mucho por hacer, pero cuidar la glucosa y la insulina quizás pueda ayudar a que futuros pacientes tengan menor probabilidad de ser detectados con diabetes (como era de imaginarse).