Comparativo de Modelos

Sebastian Jaremczuk

2020-04-17

carga de datos

COMPARAR MODELOS

Una vez que se han entrenado y optimizado distintos modelos, se tiene que identificar cuál de ellos consigue mejores resultados. La manera en la que se van a comparar los modelos es a través de métricas de Validación y el error en el Test.

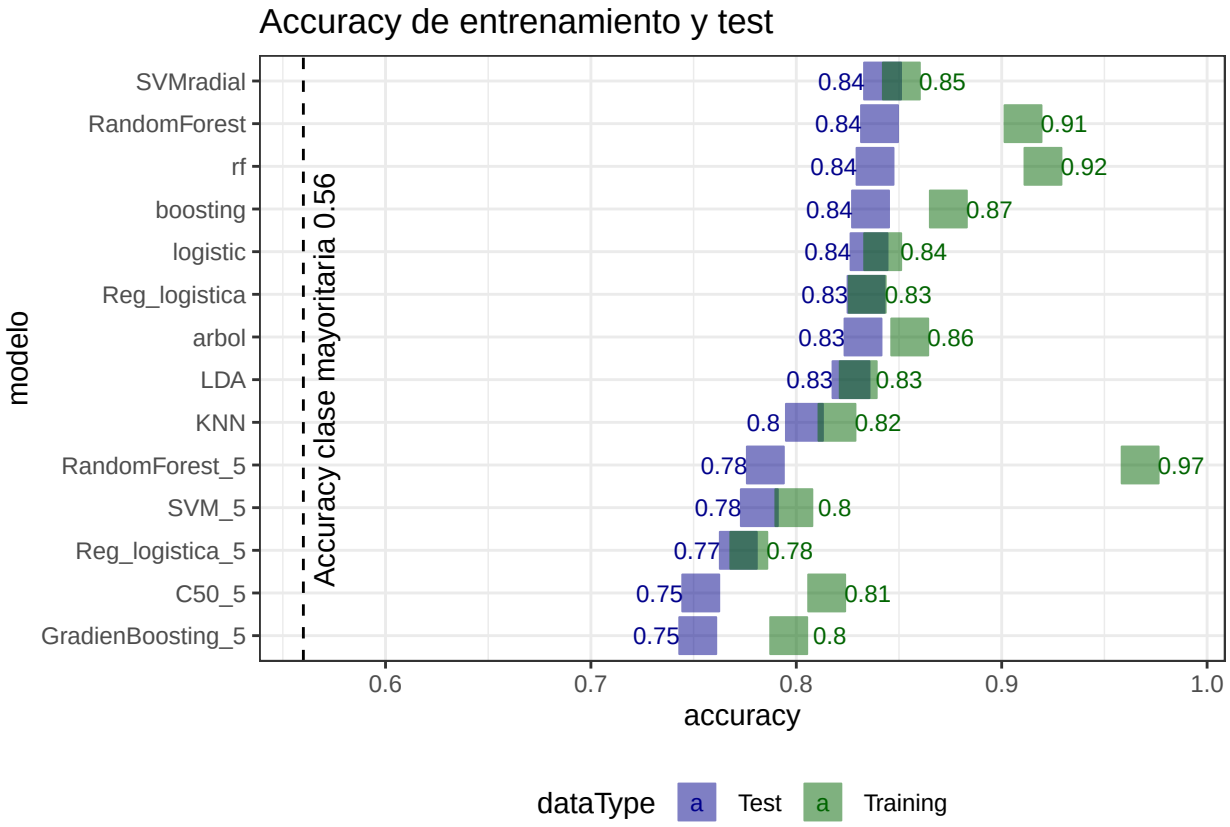
Este análisis se realiza con todos los modelos y sus variantes usando distintos datasets (los distintos métodos y para los que usan todos los predictores o una selección de los mismos **??**)

Métricas de valicación

Las métricas obtenidas mediante validación (cv, etc) son estimaciones de la capacidad que tiene un modelo al predecir nuevas observaciones.

Table 1: Resumen comparativo de algunos los modelos empleados

object	Test	Training	dataset
SVMradial	0.8419898	0.8511438	DataSet-Completo
RandomForest	0.8405267	0.9103729	DataSet-Alternativa1
rf	0.8383321	0.9200877	DataSet-Completo
boosting	0.8361375	0.8740207	DataSet-Completo
logistic	0.8354060	0.8420558	DataSet-Completo
Reg_logistica	0.8339429	0.8345346	DataSet-Alternativa1
arbol	0.8324799	0.8552178	DataSet-Completo
LDA	0.8266277	0.8301473	DataSet-Completo
KNN	0.8039503	0.8198057	DataSet-Completo
RandomForest_5	0.7849305	0.9674083	DataSet-Alternativa2
SVM_5	0.7820044	0.7988092	DataSet-Alternativa2
Reg_logistica_5	0.7717630	0.7768725	DataSet-Alternativa2
C50_5	0.7534748	0.8147916	DataSet-Alternativa2
GradienBoosting_5	0.7520117	0.7963021	DataSet-Alternativa2



Los vaores expresados en el gráfico son aquellos que una vez optimizado los parámetros con Crossvalidation, se entrena el modelo sin particiones utilizando todas las observaciones como train.

conclusión de comparacion modelos

En los modelos que se realizan con datos completos, podemos determinar que el mejor modelo entrenado es el que emplea el método SVM. También resulta ser el mejor cuando se elimina la variable que mas influencia tiene en el resultado. De los métodos explicativos se puede decir que la regresión logistica seguido del arbol simple dan buenos resultados y no muy alejado de las métricas de SVM.

Po rotro lado, los distintos modelos de Random Forest dan muy buenos resultados pero hay una diferencia muy grande en comparación a los otros métodos entre train y test, por lo que hay que ser precavido por el riesgo de sobreajuste.

Por lo tanto, si lo importante es elegir unmodelo que tenga mejor capacidad predictiva, con estas combinaciones de datos, la mejor opción es un SVM. No obstante, si se prioriza la interpretabilidad del modelo para extraer conclusiones, se podría seleccionar el modelo Regresión Logística o arbol simple.