Universidad de Buenos Aires
Facultad de ingeniería

Organización de datos (75.06)
Trabajo Práctico 2
2°Cuatrimestre 2020

María Belén Aguada
96851

Orive, María Sol
91351

TP2 - Organización de datos

TABLA 1

Nombre	**Detalle**	**Función**
Base	Se encarga de encodear todas las clases categóricas de baja cardinalidad. Además rellena con el promedio la “edad” que tiene un porcentaje de 20% de missing values.	common_preprocessing
Escalador	Se encarga de normalizar y escalar los datos para los modelos sensibles a ruido como KNN. Además llama al preprocesamiento base.	knn_preprocessing
Get Train	Se encarga de separar el dataset en una parte para train y otra (el 20%) para validación	get_train_test_data
Get Dataset	Devuelve el dataset completo	get_dataset
Friends and family	Agrega una columna donde agrupa las columnas del dataset que indicar que se fue con un acompañante (ya sea pariente o amigos)	ff_column_preprocessing
Minimal features	Elimina todas las columnas con algún porcentaje de missing values.	min_features_preprocessing
Minimal features scaler	Es como el minimal features pero normaliza los datos.	scaler_min_features_preprocessing
Friends and family scaler	Es como el friends and family pero normaliza los datos.	scaler_ff_column_preprocessing

TABLA 2

Modelo	Preprocesamiento	AUC-ROC	Accuracy	Precision	Recall	F1 score
1-DecisionTree	Base	0.7705	0.7701	0.76	0.77	0.77
1-KNN	Escalador	0.7988	0.8199	0.83	0.80	0.81
1-SVM	Escalador	0.8228	0.8509	0.88	0.82	0.84
1-Random Forest	Base	0.7818	0.8136	0.84	0.78	0.79
2-XGBoost	Friends and family scaler	0.7905	0.8074	0.81	0.79	0.80
2-Stacking	Escalador	0.8238	0.8385	0.84	0.82	0.83
1-NaiveBayes	Base	0.6331	0.6708	0.66	0.63	0.63
2-RedesNeuronales	Base	0.7771	0.7888	0.78	0.78	0.78

Conclusión

Durante la primera etapa del trabajo conseguimos un accuracy de 0.8052 mientras que en las métricas de esta etapa se llegaron a ver accuracy de hasta 0.8509 siendo una mejora de 4% .Además tomamos como métrica AUC-ROC que nos da más información sobre el comportamiento de los modelos no sólo en un % de acierto sino también sobre qué tipo de errores cometió (false positive, false negative).
Otro punto a destacar es que el accuracy obtenido en la versión anterior estaba dado por los datos mismos de “entrenamiento” mientras que en este caso dejamos un set de datos específicamente separados para poder calcular esta métrica.

Luego de probar una gran variedad de modelos, recomendamos utilizar SVM para este caso no sólo porque es el modelo que mejores métricas generales tiene sino que además es el que tiene mayor cantidad de true positive que son aquellos usuarios a los que vamos a estar atacando con las comunicaciones de marketing y un bajo número de false positive que serían esos usuarios donde vamos a comunicarles la intención de que vuelvan y nos generarían pérdida de dinero.

Además pudimos notar que modelos con mayor complejidad como Redes Neuronales o incluso Stacking donde estamos usando el poder de 3 de los modelos que acá probamos por separado no logran la misma eficiencia que un modelo más simple como SVM en este caso.

Un caso llamativo es Random Forest que suele ser muy buen clasificador, poniendo una gran cantidad de árboles a procesar y sin embargo tuvo peor resultado que la mayoría de los modelos, esto de todas maneras cumple la regla de "modelo menos complejo, tiende a tener menor varianza y generalizar mejor".