Trabajo Práctico parte 2 — FIUFIP: que no se escape nadie

[66.20] Organización de Datos
Curso 2
Primer cuatrimestre de 2021

Alumnos	Padrón	Email
ARRACHEA, Tomás	104393	tarrachea@fi.uba.ar
CAPELLI, Sebastián	98316	scapelli@fi.uba.ar

Índice

1. Tabla 1: preprocesamientos

Preprocesamientos	Explicación breve	Nombre de la función
Dummy Variables	Es una técnica de conversión de features categóricos de tipo string a formato numérico. Por cada categoría (menos una) se genera una columna que indica si la categoría está presente.	dummy_variables
Ordinal Encode	Es una técnica de conversión de features ordinales a formato numérico. Se asigna un número a cada categoría, manteniendo el orden.	ordinal_encode
Embedding	Es una técnica de reducción de dimensionalidad. Se usa el modelo de árbol de decisión para elegir los features más importantes para dicho modelo. Se entrena un árbol con todos los features y se descarta el menos importante, y así sucesivamente hasta llegar a un threshold de importancia del feature.	embedded
Features colineales	A partir del análisis exploratorio del tp1 podemos eliminar algunos features muy correlacionados, para evitar que haya colinealidad.	remove_irrelevant_features
Scaling	Es una técnica para escalar los datos, de forma que la media total valga 0 y la varianza valga 1.	escalar
Normalizar	Es una técnica para normalizar los datos, de forma que la norma de cada fila valga 1.	normalizar
PCA	Es una técnica de reducción de dimensionalidad. Proyecta los datos en una dimensión menor, tal que se maximice su varianza. Se puede elegir como parámetro el porcentaje de varianza original que se debe mantener.	pca

2. Tabla 2: modelos

	Modelo	Preprocesamiento	AUC-ROC	Accuracy	Precision	Precision	F1 score
1	arbol	Features colineales, Dummy Variables, Embedding	0.91	0,85	0,80	0,53	0,64
2	knn	Features colineales, Dummy Variables, Embedding	0,89	0,84	0,72	0,56	0,63
3	svm	Ordinal Encoding, Dummy Variables, Normalizer, PCA	0,75	0,78	0,59	0,31	0,41
4	nb	Features colineales, Dummy Variables	0,89	0,81	0,75	0,30	0,43
5	red_neuronal	Features colineales, Dummy Variables, Scaling	0,90	0,84	0,71	0,60	0,65

3. Conclusiones

A lo largo del TP fuimos entrenando distintos tipos de modelos, apendiendo más en profundidad las debilidades y fortalezas de cada uno. Para SVM, notamos que el modelo es muy costoso, por lo que solo pudimos entrenarlo para el 15 % del dataset. Si se hubiese entrenado con la misma

cantidad de datos que el resto de modelos, su performance podría haber sido mejor. Lo mismo sucedió para el modelo de KNN.

Luego del análisis realizado, podemos concluir que el modelo recomendado es el árbol de desición. El modelo obtuvo muy buena performance comparado con el resto, y cuenta con las ventajas de ser un modelo simple, interpretable y rápido para ejecutar. Por lo tanto, para este dataset resulta la mejor opción. Al compararlo con el modelo baseline, se ve que hay muchas similitudes. Los features más importantes para el modelo son aquellos que fueron seleccionados también en el baseline. El baseline era más simple y tenía un accuracy menor, pero el funcionamiento es el mismo. Si se necesitase un modelo que de una baja cantidad de falsos positivos, es decir, una alta precisión, se recomendaría usar el modelo 3 de Naive Bayes, que tiene la particularidad de ser una combinación de NB multinomial y gausiano. El modelo tiene un score de 0,96 de precision, a costa de un recall muy bajo, de 0,18. El modelo no captará bien todos los positivos, pero tendrá un error de falsos positivos muy bajo.

En cambio, si se necesitara un modelo con alto recall, se recomendaría el modelo 2 de Naive Bayes, que es el modelo multinomial. Tiene un score de 0,74 de recall, que es el más alto que se consiguió, que conlleva una baja precisión.