.UBAfiuba
FACULTAD DE INGENIERÍA

(75.06 - 95.58)
Organización de Datos

Trabajo Práctico - Parte 2

Alumnos

- Lukas Nahuel De Angelis Riva (Padrón: 103784)
- Nicolás Continanza (Padrón: 97576)

Primer Cuatrimestre 2021

Nombre Preprocesamiento	Explicación	Nombre de la función
Base	Selecciona las variables utilizadas en el baseline de la parte 1 del TP.	preprocessing_base_parte_1
StandardBase	Idem anterior, estandarizando las columnas del dataset.	standard_preprocessing_base_parte_1
Significantes **V**%	Recibe X_train, X_test y una varianza **[V]** a explicar por PCA. Devuelve X_train y X_test fiteados con PCA	preprocessing_significantes
Equilibrado	Equilibra la cantidad de muestras con alto y bajo poder adquisitivo para que haya 50% de cada una en el set de entrenamiento.	preprocessing_equilibrado
Mejores variables según Tree	Selecciona las mejores variables según el entrenamiento de un árbol.	preprocessing_mejores_por_arbol
Standard mejores variables según Tree	Idem anterior, estandarizando las columnas del dataset	standard_preprocessing_mejores_por_arbol

Modelo	Preprocesamiento	AUC-ROC	Accuracy	Precision	Recall	F1 Score
1- DecisionTreeClassifier	Equilibrado	0.9046	0.7960	0: 0.9479 1: 0.5482	0: 0.7739 1: 0.8657	0: 0.8521 1: 0.6713
2- KNeighborsClassifier	StandardBase	0.8950	0.8417	0: 0.8679 1: 0.7247	0: 0.9335 1: 0.5520	0: 0.8995 1: 0.6267
3- MultinomialNB	Significantes 90% + MinMax_Scaler[1]	0.8682	0.7592	0: 0.7592 1: 0.0000	0: 1.0000 1: 0.0000	0: 0.8631 1: 0.0000
4- SVC	Mejores variables según Tree	0.9123	0.8617	0: 0.8727 1: 0.8066	0: 0.9575 1: 0.5597	0: 0.9131 1: 0.6608
5- RandomForest	Sin preprocessing	0.9181	0.8567	0: 0.8733 1: 0.7782	0: 0.9489 1: 0.5658	0: 0.9095 1: 0.6552
6- LogisticRegression	Standard mejores variables según Tree	0.8926	0.8380	0: 0.8644 1: 0.7182	0: 0.9330 1: 0.5383	0: 0.8974 1: 0.6153
7- BaggingClassifier	Base	0.8965	0.8508	0: 0.8812	0: 0.9287	0: 0.9043

[1] MinMax_Scaler escala los datos entre 0 y 1. No está en el archivo de preprocessing.py pues sólo lo utiliza NaiveBayes.

				1: 0.7289	1: 0.6051	1: 0.6613
8- GradientBoosting	Base	0.9144	0.8565	0: 0.8733 1: 0.7777	0: 0.9487 1: 0.5658	0: 0.9094 1: 0.6551
9- AdaBoost (base estimator default)	Base	0.9182	0.8592	0: 0.8738 1: 0.7995	0: 0.9521 1: 0.5663	0: 0.9113 1: 0.6595
10- Redes Neuronales	StandardBase	0.9005	0.8456	0: 0.8812 1: 0.7089	0: 0.9207 1: 0.6087	0: 0.9005 1: 0.6550
11- Voting	Base	0.9124	0.8656	0: 0.8800 1: 0.7988	0: 0.9531 1: 0.5887	0: 0.9151 1: 0.6779

Conclusiones

En base a lo analizado en los notebooks y los resultados obtenidos en la tabla, recomendamos el uso de un modelo de 11-Voting, que se trata de un ensamble con los siguientes modelos:

- `DecisionTreeClassifier(criterion='gini', min_samples_leaf=50, max_depth=11)`
- `SVC(C=100, gamma=0.0001, probability=True)`
- `RandomForestClassifier(n_estimators=850, criterion='gini', max_depth=12, random_state=27)`
- `BaggingClassifier(n_estimators=580, random_state=27)`
- `AdaBoostClassifier(n_estimators=1501)`

Y el preprocessing equilibrado. Esta decisión fue tomada en parte porque en este ensamble colaboran algunos de los mejores modelos de la tabla, como podrían ser AdaBoost-Base (que obtuvo el mejor puntaje de RocAUC) y Bagging que sobresalió entre los de su tipo. Pero además colaboran modelos que fueron entrenados y obtuvieron resultados relativamente buenos. Además, en la tabla se puede ver que el puntaje de RocAUC del modelo recomendado es el tercer mejor, y el accuracy es el mejor de todos. Por lo que creemos que efectivamente realizará un buen trabajo para predecir.

En comparación con el baseline implementado en la parte 1 de este TP, podemos decir que la mejora fue significativa, alcanzando un accuracy del 86.6%, en comparación con el 84% obtenido inicialmente. Además, tenemos una mejora muy marcada en la predicción de casos de alto poder adquisitivo, que en la primera parte había sido muy pobre.

Queremos destacar una situación que ocurrió durante la elaboración del trabajo práctico y es la alta performance en general de los modelos basados en árboles de clasificación (como pueden ser AdaBoost, Bagging, RandomForest o un árbol mismamente). Esto nos

sorprendió, dado que estos modelos suelen ser más simples que los demás, pero aún así resultaron en una fuerte competencia.

Luego, queremos notar que en general los modelos no lograron obtener métricas altas para recall de unos (sin bajar demasiado el recall de ceros), y esta situación influyó notablemente a lo largo de todo el trabajo práctico, tomando el rol de cuello de botella para todos ellos. Véase que el único modelo en la tabla que obtuvo valores altos para esta métrica sacrificó otras métricas como podría ser el recall de ceros y el accuracy general y la precisión para los unos.

Dicho esto, si en lugar de mirar solamente la métrica AUC-ROC tuviéramos que elegir un modelo con el solo objetivo de minimizar la cantidad de falsos positivos, la recomendación sería usar el modelo 4- SVC entrenado con el preprocesamiento mejores variables según Tree y con los siguientes hiperparámetros:

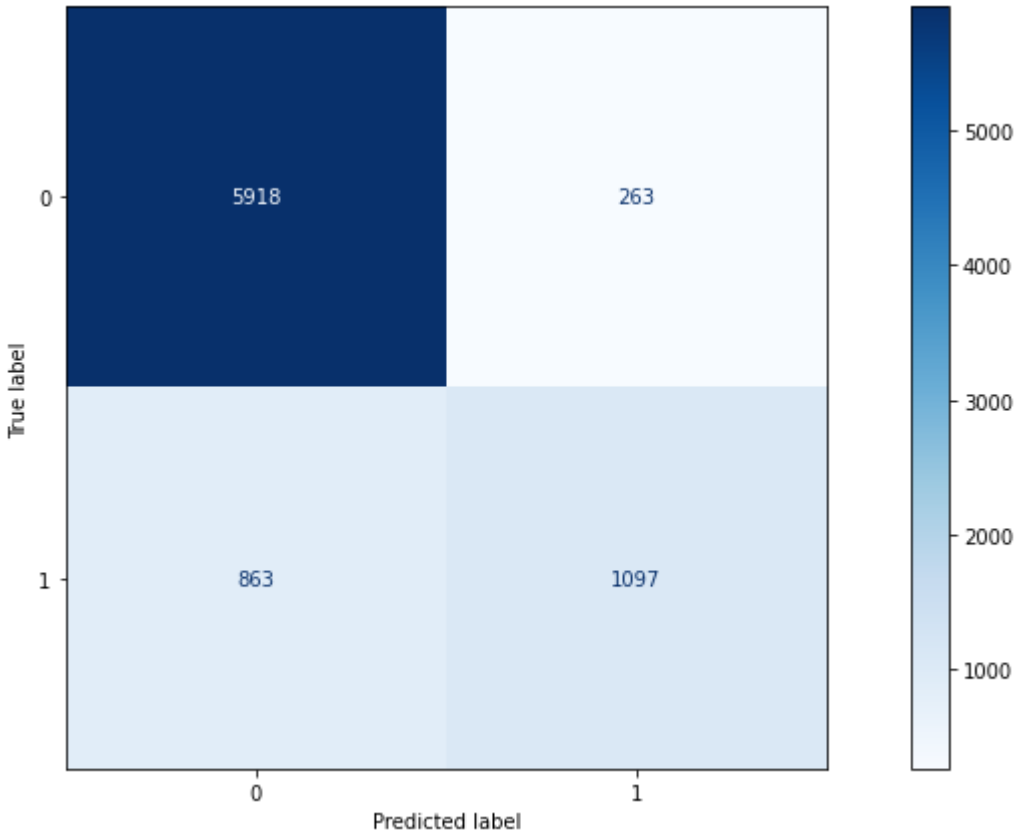
- `C = 100.0`
- `gamma = 0.0001`
- `kernel = 'rbf'`

Cuyas métricas fueron las siguientes:

Modelo	Preprocesamiento	AUC-ROC	Accuracy	Precision	Recall	F1 Score
4- SVC	Mejores variables según Tree	0.9123	0.8617	0: 0.8727 1: 0.8066	0: 0.9575 1: 0.5597	0: 0.9131 1: 0.6608

Vemos aquí que la precisión de unos es alta y su recall es el común para los modelos que entrenamos en el trabajo.

Podemos ver ahora la matriz de confusión para corroborar lo pedido:



Y vemos que la esquina superior derecha tiene pocas instancias.

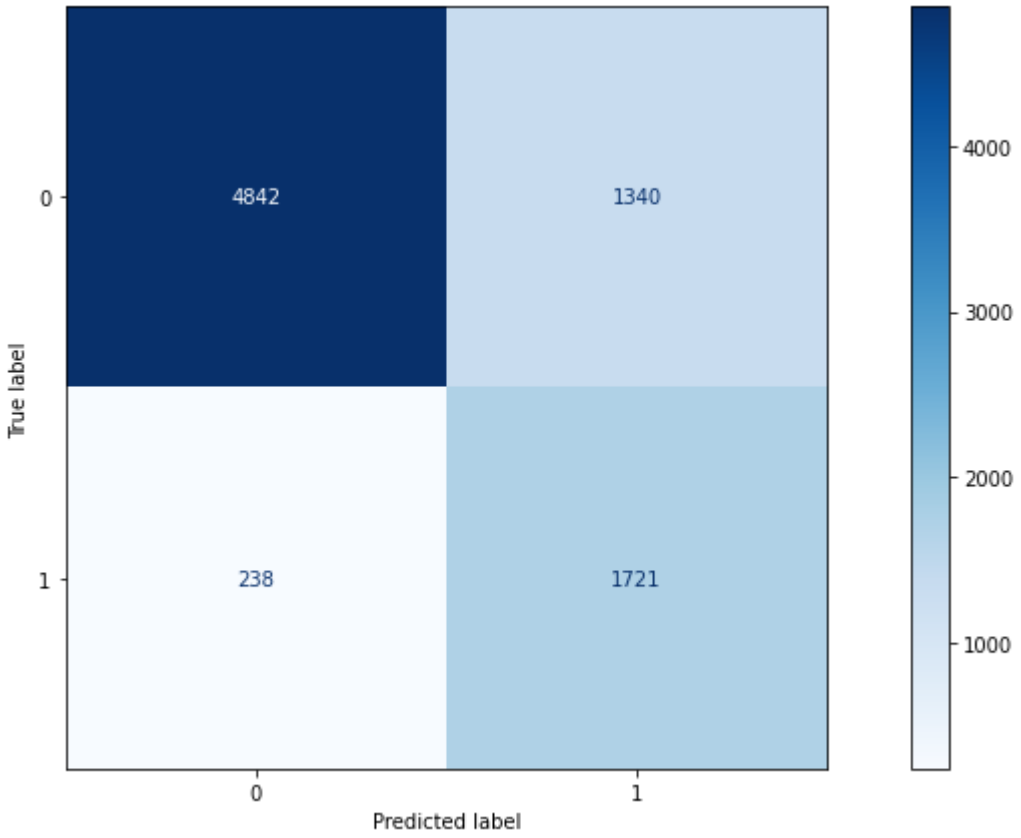
Si en cambio quisiéramos obtener una lista de aquellos individuos con un potencial poder adquisitivo alto, sin importar la cantidad de falsos positivos que allí encontremos, recomendaríamos usar un RandomForest entrenado con el preprocessing equilibrado con los siguientes hiperparámetros:

- `n_estimators = 1001`
- `criterion = 'entropy'`
- `max_depth = 15`
- `max_features = 8`

Donde obtuvimos las siguientes métricas:

Modelo	Preprocesamiento	AUC-ROC	Accuracy	Precision	Recall	F1 Score
RandomForest	Equilibrado	0.9141	0.8062	0: 0.9531 1: 0.5622	0: 0.7832 1: 0.8785	0: 0.8599 1: 0.6857

Nuevamente, para comprobar lo pedido presentamos la matriz de confusión para dicho modelo:



Vemos que la celda inferior izquierda posee pocas instancias y que la diagonal de la matriz posee la gran parte de las instancias (eso significa que sigue haciendo un buen trabajo de predicción).

Comparación parte 1 vs parte 2

Observamos una mejora notable de las métricas para los modelos de la parte 2 que, dicho sea de paso, fueron testeados sobre datos de test, mientras que en la parte 1 no.

Las métricas obtenidas para la primera parte del trabajo son:

Modelo	Preprocesamiento	Accuracy	Precision	Recall	F1 Score
Baseline	Sin preprocessing	0.8446	0: 0.8587 1: 0.7695	0: 0.9519 1: 0.5062	0: 0.9029 1: 0.6107

Vemos que nuestro clasificador supera por 2% de accuracy a este modelo, y además todas las demás métricas son superiores.

De todas formas nos sorprende que nuestro modelo que nació de la simple observación de los datos mediante gráficos aún así supere a gran parte de los modelos realizados en la elaboración de la parte 2.