Lucas Ruiz Zapata

Preguntas sobre el modelo

• ¿Cuál es el problema de clasificación que están tratando de resolver?

Estamos intentando predecir si un paciente de covid19 es de alto riesgo o no

• ¿Qué modelo de clasificación seleccionaron como el mejor y por qué?

El mejor modelo de clasificación es lightgbm

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	1.6711	3.2986	1.8162	0.0766	0.3287	0.4219	3.6460
gbr	Gradient Boosting Regressor	1.6794	3.3102	1.8194	0.0734	0.3293	0.4242	31.8760
ridge	Ridge Regression	1.6922	3.3505	1.8304	0.0621	0.3312	0.4276	1.0240
lar	Least Angle Regression	1.6922	3.3505	1.8304	0.0621	0.3312	0.4276	1.0120
br	Bayesian Ridge	1.6922	3.3505	1.8304	0.0621	0.3312	0.4276	1.2220
lr	Linear Regression	1.6923	3.3505	1.8304	0.0621	0.3312	0.4277	1.2420
huber	Huber Regressor	1.6462	3.3801	1.8385	0.0538	0.3344	0.4269	8.3020
ada	AdaBoost Regressor	1.7886	3.4589	1.8598	0.0317	0.3306	0.4291	15.6700
omp	Orthogonal Matching Pursuit	1.7568	3.4905	1.8683	0.0229	0.3380	0.4435	1.0060
en	Elastic Net	1.7608	3.4910	1.8684	0.0228	0.3381	0.4445	1.0420
lasso	Lasso Regression	1.7649	3.4941	1.8693	0.0219	0.3383	0.4455	1.0420
llar	Lasso Least Angle Regression	1.7649	3.4941	1.8693	0.0219	0.3383	0.4455	1.0100
rf	Random Forest Regressor	1.6765	3.5050	1.8722	0.0188	0.3375	0.4225	97.5580
dummy	Dummy Regressor	1.7973	3.5724	1.8901	-0.0000	0.3418	0.4536	1.0560
et	Extra Trees Regressor	1.6811	3.6352	1.9066	-0.0176	0.3438	0.4236	106.1900
dt	Decision Tree Regressor	1.6878	3.7718	1.9421	-0.0558	0.3505	0.4249	2.2820
knn	K Neighbors Regressor	1.7064	3.9536	1.9884	-0.1067	0.3534	0.4255	36.8580
par	Passive Aggressive Regressor	2.1652	6.4636	2.4519	-0.8078	0.4606	0.4259	1.7820

Esto se debe a que ha cumplido casi todas las pruebas mejor que el resto de los modelos.

Él único que lo supera en una prueba es el “huber” que lo supera en MAE con 1.6462

Lucas Ruiz Zapata

• ¿Cómo evaluaron el rendimiento del modelo en el conjunto de pruebas? ¿Cuáles fueron los resultados?

El modelo fue comparado con el resto con la biblioteca de pycaret y en concreto usamos la función “compare_models()” para comparar todos los modelos a las mismas pruebas y seleccionar el que de mejores resultado de estas.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	1.6711	3.2986	1.8162	0.0766	0.3287	0.4219	3.6460
gbr	Gradient Boosting Regressor	1.6794	3.3102	1.8194	0.0734	0.3293	0.4242	31.8760
ridge	Ridge Regression	1.6922	3.3505	1.8304	0.0621	0.3312	0.4276	1.0240
lar	Least Angle Regression	1.6922	3.3505	1.8304	0.0621	0.3312	0.4276	1.0120
br	Bayesian Ridge	1.6922	3.3505	1.8304	0.0621	0.3312	0.4276	1.2220
lr	Linear Regression	1.6923	3.3505	1.8304	0.0621	0.3312	0.4277	1.2420
huber	Huber Regressor	1.6462	3.3801	1.8385	0.0538	0.3344	0.4269	8.3020
ada	AdaBoost Regressor	1.7886	3.4589	1.8598	0.0317	0.3306	0.4291	15.6700
omp	Orthogonal Matching Pursuit	1.7568	3.4905	1.8683	0.0229	0.3380	0.4435	1.0060
en	Elastic Net	1.7608	3.4910	1.8684	0.0228	0.3381	0.4445	1.0420
lasso	Lasso Regression	1.7649	3.4941	1.8693	0.0219	0.3383	0.4455	1.0420
llar	Lasso Least Angle Regression	1.7649	3.4941	1.8693	0.0219	0.3383	0.4455	1.0100
rf	Random Forest Regressor	1.6765	3.5050	1.8722	0.0188	0.3375	0.4225	97.5580
dummy	Dummy Regressor	1.7973	3.5724	1.8901	-0.0000	0.3418	0.4536	1.0560
et	Extra Trees Regressor	1.6811	3.6352	1.9066	-0.0176	0.3438	0.4236	106.1900
dt	Decision Tree Regressor	1.6878	3.7718	1.9421	-0.0558	0.3505	0.4249	2.2820
knn	K Neighbors Regressor	1.7064	3.9536	1.9884	-0.1067	0.3534	0.4255	36.8580
par	Passive Aggressive Regressor	2.1652	6.4636	2.4519	-0.8078	0.4606	0.4259	1.7820

• ¿Cuáles fueron las métricas de rendimiento que utilizaron para evaluar el modelo?

Las métricas de rendimiento que hemos utilizado para evaluar el modelo son las siguientes

- **MAE**: El error absoluto medio (MAE) es la media de las diferencias absolutas entre los valores predichos y los valores reales. Es una medida de la magnitud del error sin tener en cuenta su dirección. Cuanto menor sea el valor de MAE, mejor será el modelo.
- **MSE**: El error cuadrático medio (MSE) es la media de los errores al cuadrado entre los valores predichos y los valores reales. Es una medida de la magnitud del error sin tener en cuenta su dirección. MSE es más sensible a los valores atípicos que MAE.
- **RMSE**: La raíz del error cuadrático medio (RMSE) es la raíz cuadrada de MSE. Es una medida de la magnitud del error sin tener en cuenta su dirección. RMSE es más sensible a los valores atípicos que MAE.
- **R2**: El coeficiente de determinación (R2) es una medida de la proporción de la varianza en la variable dependiente que se explica por la variable independiente(s) en un

modelo de regresión. R2 varía entre 0 y 1, y cuanto más cercano sea a 1, mejor será el modelo.

- **RMSLE**: La raíz del error cuadrático medio de los registros (RMSLE) es la raíz cuadrada de la media de los errores al cuadrado de los registros entre los valores predichos y los valores reales. Es una medida de la magnitud del error sin tener en cuenta su dirección. RMSLE se utiliza a menudo en problemas de regresión donde las variables objetivo tienen una distribución sesgada.
- **MAPE**: El error porcentual absoluto medio (MAPE) es la media de los errores porcentuales absolutos entre los valores predichos y los valores reales. Es una medida de la magnitud del error sin tener en cuenta su dirección. MAPE se utiliza a menudo en problemas de pronóstico.

• ¿Cuáles fueron los hiperparámetros que utilizaron para mejorar el rendimiento del modelo?

Hemos utilizado las mismas métricas para comparar el resultado del entrenamiento del modelo con los datos

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	1.6319	3.2133	1.7926	0.0943	0.3243	0.4116
1	1.6305	3.2127	1.7924	0.0937	0.3247	0.4124
2	1.6300	3.2132	1.7925	0.0906	0.3243	0.4112
3	1.6308	3.2078	1.7910	0.0965	0.3241	0.4118
4	1.6313	3.2138	1.7927	0.0952	0.3246	0.4123
5	1.6309	3.2140	1.7928	0.0902	0.3241	0.4107
6	1.6275	3.1993	1.7887	0.0911	0.3227	0.4079
7	1.6275	3.2095	1.7915	0.0902	0.3241	0.4105
8	1.6303	3.2091	1.7914	0.0948	0.3238	0.4106
9	1.6294	3.2004	1.7890	0.0920	0.3225	0.4073
Mean	1.6300	3.2093	1.7915	0.0929	0.3239	0.4106
Std	0.0014	0.0052	0.0014	0.0022	0.0007	0.0017

Lucas Ruiz Zapata

• ¿Cómo visualizaron los resultados del modelo? ¿Qué información pueden extraer de la visualización?

Hemos usado la función ``` predict_model() ``` para hacer una predicción usando los datos obtenidos y el conjunto de datos de prueba.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	1.6684	3.2892	1.8136	0.0790	0.3282	0.4211

	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	PNEUMONIA	AGE	\
0	0	1	Female	hospitalized	1.0	65.0	
1	0	1	Male	hospitalized	1.0	72.0	
2	0	1	Male	not hospitalized	0.0	55.0	
3	0	1	Female	hospitalized	0.0	53.0	
4	0	1	Male	hospitalized	0.0	68.0	
...	...	...	...	...	...	...	
1048570	0	13	Male	hospitalized	0.0	40.0	
1048571	1	13	Male	not hospitalized	0.0	51.0	
1048572	0	13	Male	hospitalized	0.0	55.0	
1048573	0	13	Male	hospitalized	0.0	28.0	
1048574	0	13	Male	hospitalized	0.0	52.0	

	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	\
0	0.0	0.0	0.0	0.0	0.0	1.0	
1	0.0	0.0	0.0	0.0	0.0	1.0	
2	0.0	1.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	1.0	0.0	0.0	0.0	1.0	
...	...	...	...	...	...	...	
1048570	0.0	0.0	0.0	0.0	0.0	0.0	
1048571	0.0	0.0	0.0	0.0	0.0	1.0	
1048572	0.0	0.0	0.0	0.0	0.0	0.0	
1048573	0.0	0.0	0.0	0.0	0.0	0.0	
1048574	0.0	0.0	0.0	0.0	0.0	0.0	

...

1048573	0.0	7	5.648801
1048574	0.0	7	5.543871

[1032572 rows x 20 columns]

• ¿Cómo interpretaron el modelo? ¿Qué características o variable son más relevantes en la toma de decisiones del modelo?

Según las comprobaciones de los datos, el modelo usa las variables de AGE y MEDICAL_UNIT como las más relevantes

Lucas Ruiz Zapata