Universidad Autónoma de Nuevo León

Maestría en Ciencia de Datos

Aprendizaje Automatico

Producto Integrador – PIA: Modelo de clasificación, con validación cruzada (criterio ROC AUC)

Profesor: José Anastacio Hernández Saldaña

Alumno: Sergio Bernal Cortez

Matricula: 1581391

San Nicolás de los Garza, Nuevo León a 28 de julio de 2024

Índice

Objetivos

Objetivo General

Hacer uso de las técnicas para la clasificación de clases de un conjunto de datos

Objetivo Especifico

Se hará uso de una base de datos para emplear un modelo de clasificación con que tenga una calificación de al menos 0.75, mediante el criterio Receiver Operating Characteristic (ROC) y el puntaje será determinado por el área bajo la curva (AUC)

Introducción

En este reporte se verán algunos ejemplos para clasificar datos, los cuales vienen dados por el archivo “train.csv” el cual hace alusión a documentos y contiene un total de 9 variables con 9239 registros, de las cuales 8 son valores numéricos o bien, propiedades de cada documento, mientras que una (la novena) hace alusión a si hay o no un compromiso o entusiasmo por parte quizá de los lectores de cada documento.

Resultados

Lo primero que se hizo con los datos fue revisar de que tipo eran, resultando en que todos excepto 1 eran valores numéricos, es decir, había una variable que de entrada ya segregaba a nuestros datos, por lo que, para efectos del análisis de clasificación, esta fue nuestra variable dependiente (y) y el resto de las variables fueron las independientes (x), adicional, se contaba con el dato de identificación de cada documento, el cual, a pesar de ser un número, no se considero como variable y fue eliminado de los datos

De manera general, el conjunto de datos quedó de la siguiente manera

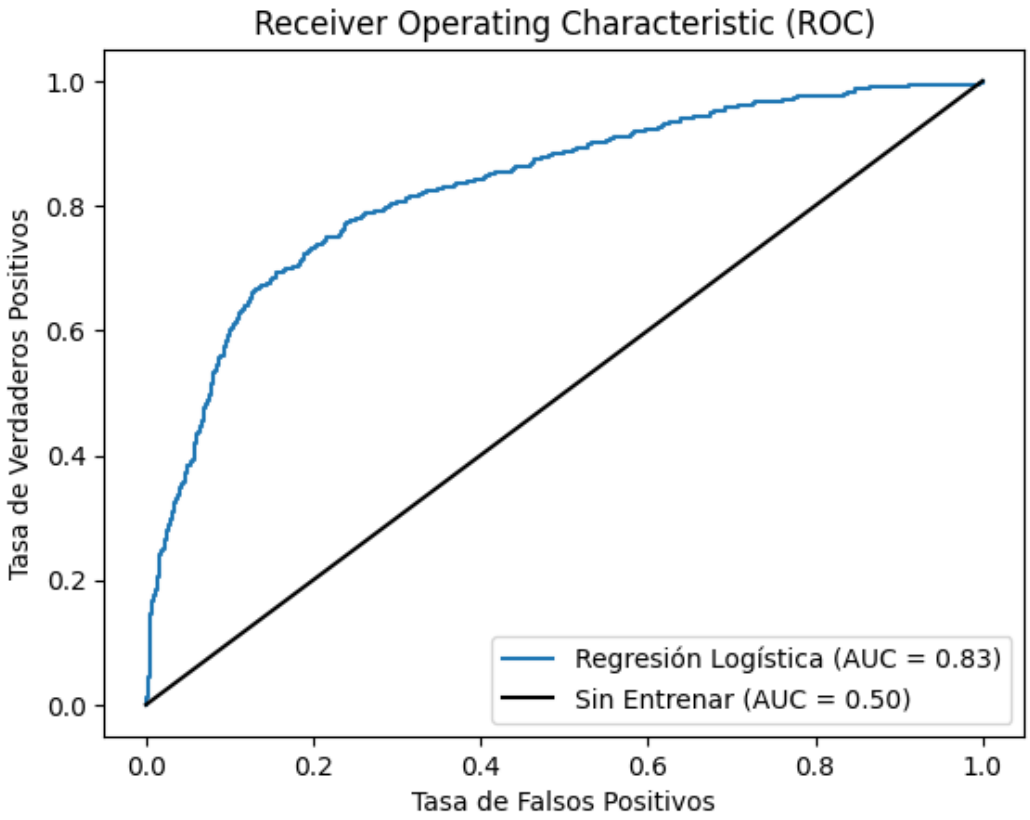
	title_word_count	document_entropy	freshness	easiness	fraction_stopword_presence	normalization_rate	speaker_speed	silent_period_rate	engagement
0	9	7.753995	16310	75.583936	0.553664	0.034049	2.997753	0.000000	1
1	6	8.305269	15410	86.870523	0.584498	0.018763	2.635789	0.000000	0
2	3	7.965583	15680	81.915968	0.605685	0.030720	2.538095	0.000000	0
3	9	8.142877	15610	80.148937	0.593664	0.016873	2.259055	0.000000	0
4	9	8.161250	14920	76.907549	0.581637	0.023412	2.420000	0.000000	0
...	...	...	...	...	...	...	...	...	...
9234	7	7.820262	14170	82.302473	0.587838	0.027449	2.120000	0.250322	0
9235	6	8.781639	14410	83.361440	0.588235	0.012465	2.251447	0.000000	0
9236	5	7.643789	16240	82.425897	0.611600	0.034213	2.728182	0.000000	0
9237	6	7.985593	14500	87.565381	0.630815	0.020167	2.529861	0.300094	0
9238	7	6.593383	15900	94.200932	0.647826	0.004348	2.775000	0.049792	1

9239 rows × 9 columns

Regresión Logística

Lo primero que se hizo con los datos fue una clasificación únicamente mediante regresión logística, el cual obtuvo una calificación de **0.83**, una muy buena sin necesidad de compararlo con otros métodos.

El resultado fue bueno, y fue un muy buen inicio para el análisis sin embargo el fin del ejercicio es obtener el mejor método después de ponerlos a competir entre si.

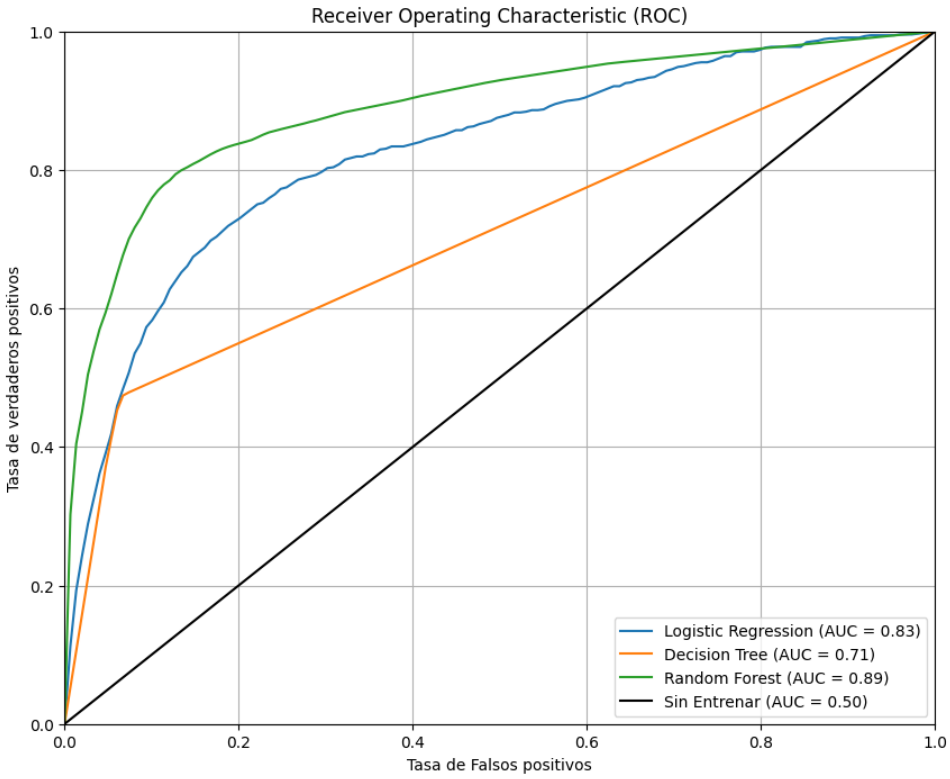


Comparación entre varios métodos

En esta segunda parte, utilicé 3 modelos de clasificación, incluyendo el anterior más otros dos nuevos, los cuales fueron los siguientes, seguidos de sus puntuaciones. (AUC)

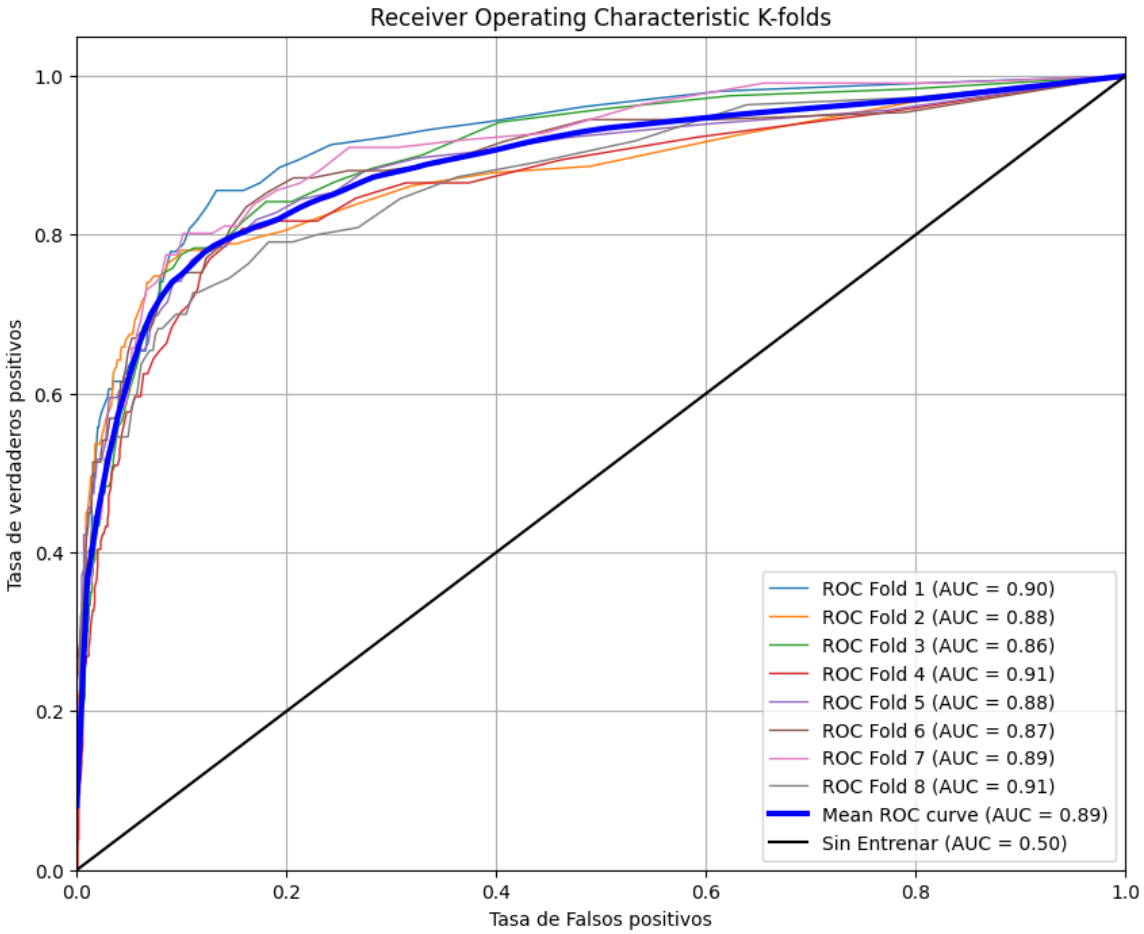
Clasificador	AUC
Logistic regression	0.83
Decision Tree	0.71
Random Forest	0.89

Con los resultados de la tabla podemos que, a pesar de ya haber tenido un buen resultado con logistic regression, random forest fue mejor, parece que no por mucho, pero en cuestión de los modelos que se utilizaron me fue algo complicado ver de que manera subir las puntuaciones, es decir, a pesar de que en cuestión de escala no parezca demasiada diferencia, si es considerable la mejora de este último modelo versus el primero.



KFold

Finalmente, el tercer método fue KFold, empleando 8 particiones donde todos los resultados fueron superiores a **0.80**, siendo el mayor **0.91** (dos veces obtenido) y un promedio de **0.89**, de nuevo se logró mejorar un poco el resultado obtenido previamente, sin embargo en comparación con el promedio fue prácticamente igual de bueno que el anterior.



Conclusión

Para realizar estos análisis, revisé el detalle de los modelos de selección de sklearn, tuve preocupación de que los datos fueran a ser más engañosos o de mayor dificultad que los ejemplos que se usan para practicar con los modelos, afortunadamente desde el inicio tuve buenos resultado y quise enriquecer más el análisis con más ejemplos y complementar este trabajo, fueron buenos resultados en general.