UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

PIA

Modelos de Clasificación

Materia: MAA(Aprendizaje Automático)
Profesor: José Anastacio Hernández Saldaña
Grupo: 03
Alumno: José de Jesús Almanza Trejo

Contenido

Introducción

A continuación se presenta una serie de modelos aplicados a un set de información referente a natural language processing model (se deduce por el nombre de las features), donde la característica referente a engagement es la variable objetivo (una variable binaria).

Se realizará la búsqueda de un mejor modelo aplicando modelos de clasificación (binaria) al conjunto de datos tal que el valor del ROC AUC (Area Under de ROC Curve) se maximice, utilizando una herramienta (optuna) de búsqueda óptima de parámetros para los diferentes algoritmos.

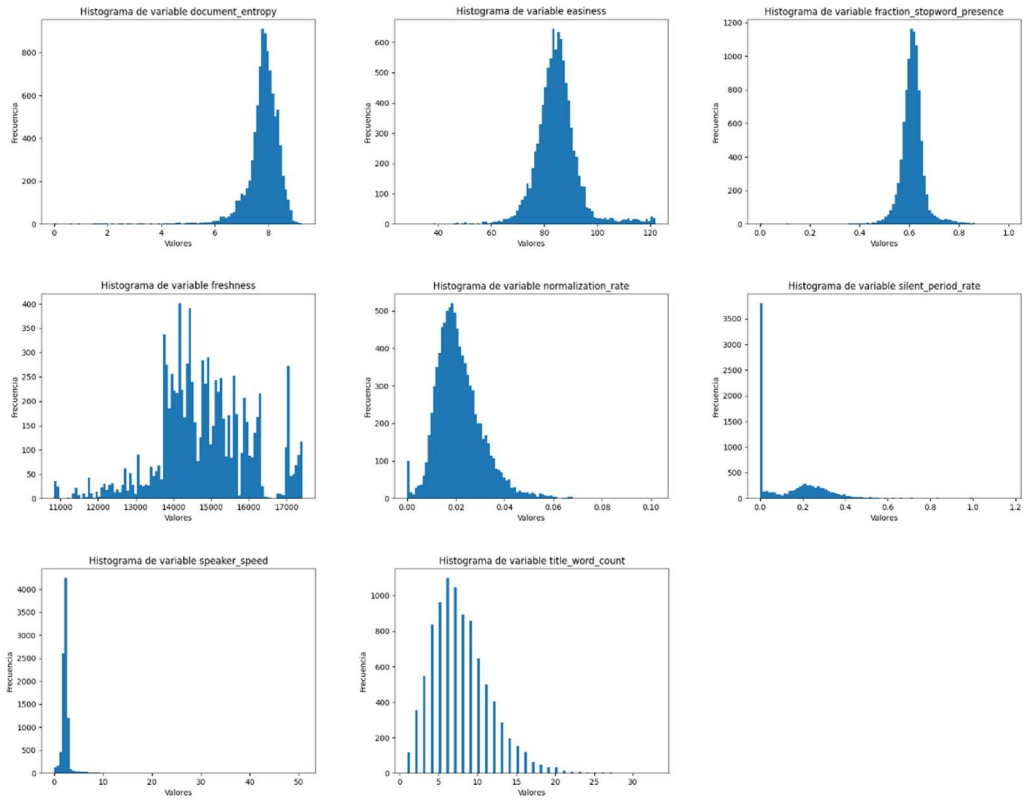
Una vez encontrado el mejor modelo se presentarán su algoritmo y parámetros.

Análisis descriptivo

Tenemos las siguientes características de las variables en el dataset:

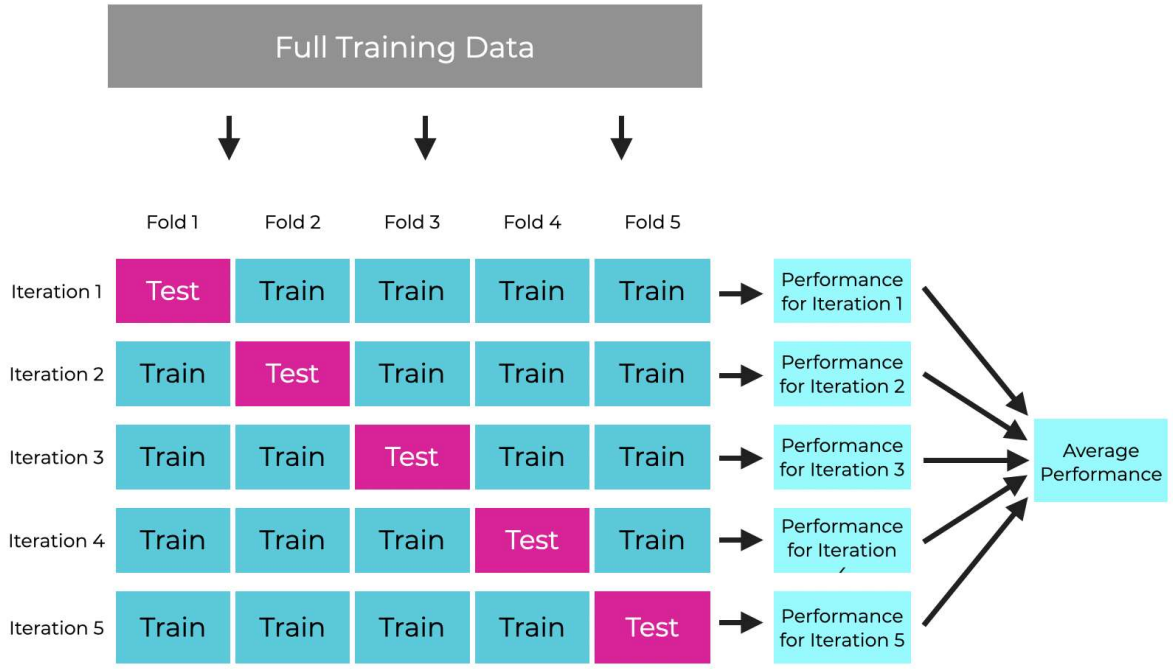
Variable	Conteo	Promedio	Mínimo	Percentil 25%	Percentil 50%	Percentil 75%	Máximo
title_word_count	9,239	7.70	1.00	5.00	7.00	10.00	33.00
document_entropy	9,239	7.79	0.00	7.59	7.88	8.16	9.28
freshness	9,239	14,808.59	10,830.00	14,070.00	14,750.00	15,600.00	17,430.00
easiness	9,239	84.76	28.21	80.41	84.48	88.39	122.03
fraction_stopword_presence	9,239	0.61	0.00	0.59	0.61	0.63	1.00
normalization_rate	9,239	0.02	0.00	0.01	0.02	0.03	0.10
speaker_speed	9,239	2.41	0.00	1.98	2.27	2.54	50.85
silent_period_rate	9,239	0.15	0.00	0.00	0.10	0.25	1.17

Se muestra el histograma de las variables:



Clasificación

Se realizan varios ejercicios de regresión con diversos algoritmos vistos en clase así como otros tipos de algoritmos de clasificación que tiene la librería sklearn (Sklearn, 2024), el modelado consistirá de la aplicación de un algoritmo de validación cruzada a partir de 10 particiones estratificadas de la base de entrenamiento en las cuales se correrá el algoritmo seleccionado con sus parámetros tomando cada una de las particiones como set de pruebas y el resto de las particiones como set de entrenamiento, se calcula el score ROC AUC para cada iteración y al final se promedia el score para mostrar el desempeño promedio del modelo.



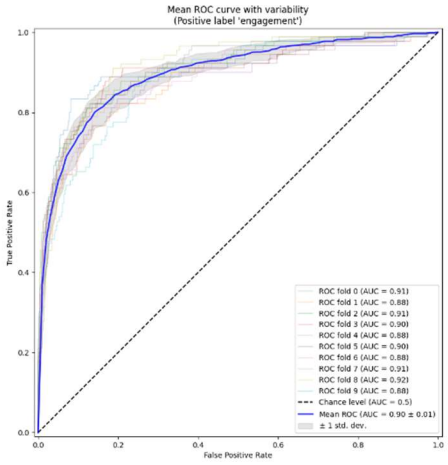
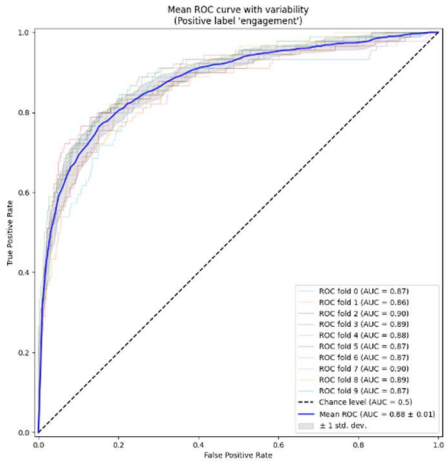
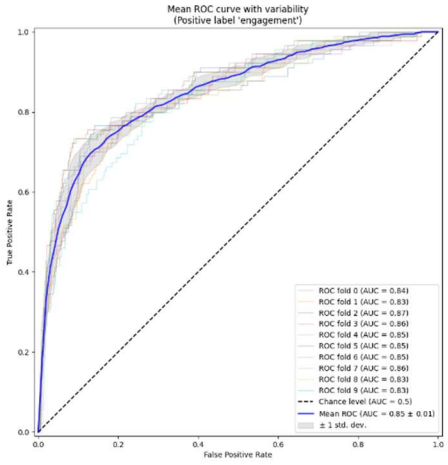
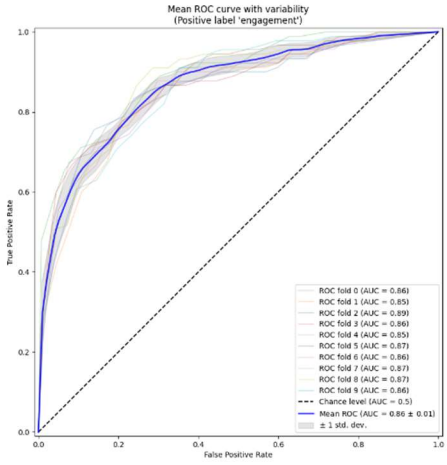
Resultados

Se corre un algoritmo de búsqueda optimizada de parámetros por medio de la librería optuna (Optuna, 2018) en donde se buscan los mejores parámetros que maximicen nuestro score seleccionado (ROC AUC) obteniendo los siguientes resultados:

Algoritmo	Parámetros	ROC AUC promedio
Logistic Regression	{'penalty': 'l2', 'solver': 'liblinear', 'C': 0.001, 'l1_ratio': None}	0.8476
Decision Tree Classifier	{'max_depth': 44, 'criterion': 'log_loss', 'min_samples_split': 0.024854806445273624, 'min_samples_leaf': 0.026777469908052437}	0.8642
Random Forest Classifier	{'n_estimators': 122, 'max_depth': 179, 'criterion': 'entropy', 'min_samples_split': 0.010022356623229676, 'min_samples_leaf': 0.01014036579990787, 'max_leaf_nodes': 42}	0.8791
Gradient Boosting Classifier	**{'learning_rate': 0.21953581594646793, 'n_estimators': 57, 'max_depth': 26, 'subsample': 0.978232013339438, 'min_samples_split': 0.2641324973039342, 'min_samples_leaf': 0.04570519683574234, 'max_leaf_nodes': 20}**	**0.8978**

El mejor modelo seleccionado es un Gradient Boosting Classifier con random_seed=123 y arroja un score promedio de ROC AUC de 0.8978.

Se muestran los gráficos de ROC AUC de los algoritmos:

Gradient Boosting	Random Forest
	
Regresión Logística	**Decision Tree**
	

Bibliografía

Optuna. (2018). Obtenido de https://optuna.readthedocs.io/en/stable/tutorial/index.html

Sklearn. (2024). Obtenido de https://scikit-learn.org/stable/supervised_learning.html