Universidad Autónoma de Nuevo León

Facultad de ciencias fisico matematicas

Maestría en ciencia de datos

Aprendizaje Automatizado

Proyecto:

Método clasificación

Catedrático: M.C. Jose Anastacio Hernandez Saldaña

Equipo : Cynthia Selene Martinez Espinoza

Matrícula : 1011238

San Nicolás de los Garza, Nuevo León, a 28 de Julio del 2024

Índice

Objetivo

El objetivo de este proyecto es construir un modelo de clasificación para predecir la probabilidad de que se otorgue un contrato con la variable objetivo`engagement` utilizando varios algoritmos de clasificación.

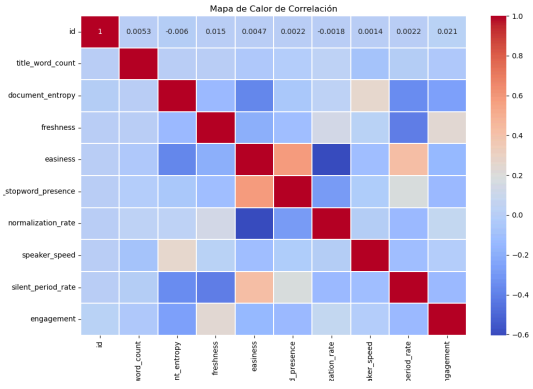
Los datos proporcionados incluyen diversas características de entrada, y se utilizaron técnicas de validación cruzada para evaluar y comparar los modelos.

Datos tomados de: Train.csv / Test.csv (Proporcionados por el catedrático de la materia)

Exploración de Datos

Se realizó un análisis descriptivo inicial de los datos para entender mejor la distribución y las correlaciones entre las características.

Revisamos la correlación de las variables , para elegir las variables a considerar.



Elegimos las variables a utilizar,

- document_entropy
- freshness
- fraction_stopword_presence
- normalization_rate

Preparación de datos

Para mantener la proporción de las clases, Normalización de características utilizando `***StandardScaler***`. - División de los datos en conjuntos de entrenamiento (80%) y validación (20%) utilizando una división segmentada.

Resultados evaluación cruzada

Se evaluaron los modelos de KNN, SVM Árbol de decisión, con validación cruzada.

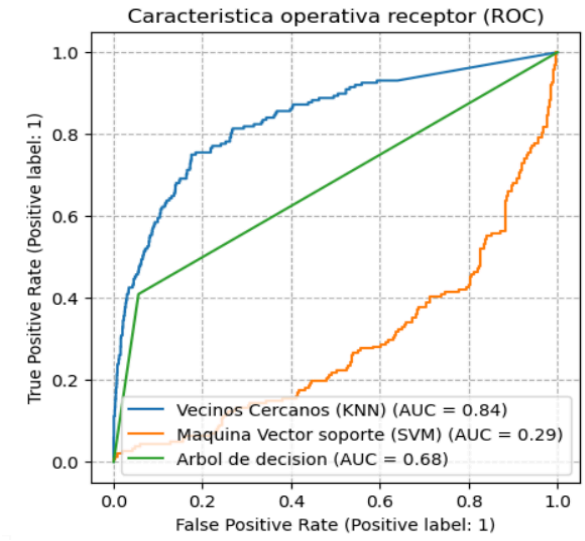
	Validación Cruzada ROC AUC
Vecinos Cercanos (KNN)	0.85
Maquina Vector soporte (SVM)	0.8
Árbol de decisión	0.71

Evaluar el modelo en el conjunto de prueba, validación.

Se buscaron los mejores parámetros para ajustar la ejecución del modelo considerando el KNN para mejor el ROC AUC. Considerando los parámetros metric='euclidean', n_neighbors=30, weights='distance'.

	Validación Accuracy	Prueba Accuracy	ROC AUC
Vecinos Cercanos (KNN)	0.92803	0.913961	0.838243
Maquina Vector soporte (SVM)	0.902597	0.898268	0.285545
Árbol de decisión	0.895022	0.887446	0.670854

Gráficas



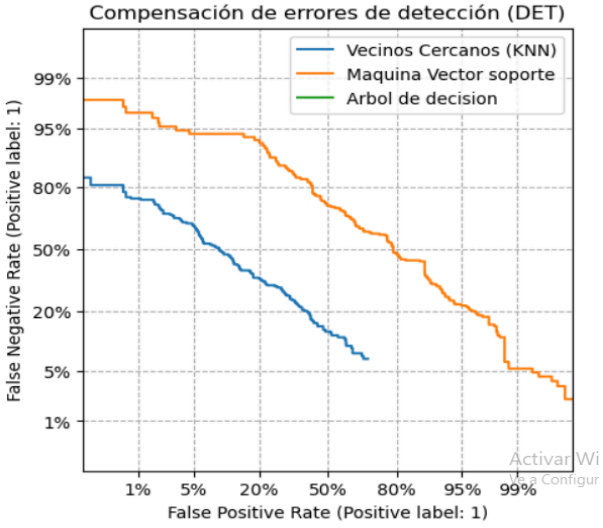
ROC nos indica que tan bien estamos adivinando

Área bajo la curva (AUC) Cuanto más grande es este número, mejor es la estrategia.

K-Nearest Neighbors (AUC = 0.84): lo que significa que es bastante buena para adivinar.

Support Vector Machine (AUC = 0.29): lo que significa que no es muy buena para adivinar.

Decision Tree (AUC = 0.68): lo que significa que es mejor que adivinar al azar, pero no tan buena como K-Nearest Neighbors.



Compensación de Errores de Detección Indica los tipo de errores cometen nuestras estrategias cuando intentan adivinar.

Cuanto más baja es la línea en esta gráfica, mejor es la estrategia porque significa que está cometiendo menos errores.

La gráfica muestra que el modelo de KNN que comete menos errores comparados con los otros modelos evaluados.

Conclusiones

El modelo K-Nearest Neighbors (KNN) fue evaluado para predecir la probabilidad de que se otorgue un contrato, basándose en el conjunto de datos proporcionado.

El rendimiento del modelo se midió utilizando la métrica de ROC AUC, la cual alcanzó un valor de 0.84. El valor de ROC AUC de 0.84 indica que el modelo tiene una buena capacidad para distinguir entre las clases de 'contrato otorgado' y 'contrato no otorgado'.

Un ROC AUC cercano a 1.0 sugiere un modelo con excelente capacidad de discriminación, mientras que un ROC AUC de 0.5 indicaría un rendimiento similar al azar. Por lo tanto, un ROC AUC de 0.84 es un resultado sólido que muestra que el modelo KNN es efectivo para el objetivo planteado.

Referencias:

Apoyo sobre los Materiales compartidos en clase., por el catedrático.

Scikit-learn documentation: https://scikit-learn.org/