Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos II

Grupo 101

22 de septiembre de 2024

Clasificación de email de spam: pre-procesamiento y baselines

Autor:

Catherine Johanna Rojas Mendoza

A01798149

Doctor:

Alfredo Esquivel Jaramillo

Clasificación de Correos Electrónicos Spam vs No Spam utilizando Modelos de Aprendizaje Automático

Se implementó un flujo de trabajo para clasificar correos electrónicos en spam y no spam utilizando modelos de aprendizaje automático. Inicialmente, los datos de correos electrónicos son cargados y procesados para extraer los cuerpos de los mensajes. Luego, se aplican técnicas de tokenización, eliminación de stop-words y limpieza de texto. Se utiliza un modelo de bag-of-words para representar numéricamente los correos electrónicos y, posteriormente, se entrenan varios clasificadores, como regresión logística, máquinas de soporte vectorial (SVM), random forests y gradient boosting. Además, se optimizan hiperparámetros mediante validación cruzada y se evalúan los modelos utilizando métricas de precisión y la curva ROC. Finalmente, el modelo entrenado se utiliza para predecir etiquetas de spam y no spam en el conjunto de prueba.

Desempeño de los clasificadores entrenados

1. Regresión Logística (LogisticRegression de *sklearn.linear_model*)

Este modelo de clasificación lineal estima la probabilidad de pertenecer a una clase particular (spam o no spam) en función de una combinación lineal de las características de entrada. Es útil cuando los datos son linealmente separables y proporciona una interpretación probabilística de las predicciones.

2. Máquinas de Soporte Vectorial (SVC de *sklearn.svm*)

Es un clasificador que busca encontrar un hiperplano óptimo que separe las clases (spam y no spam) con el mayor margen posible. SVM puede ser lineal o no lineal y es útil cuando las clases no son fácilmente separables.

3. Random Forest (RandomForestClassifier de *sklearn.ensemble*)

Es un conjunto de árboles de decisión que funcionan como clasificadores individuales. Cada árbol se entrena en una muestra aleatoria del conjunto de datos, y las predicciones se basan en el voto mayoritario de todos los árboles, lo que reduce el riesgo de sobreajuste y mejora la precisión.

4. Gradient Boosting (GradientBoostingClassifier de *sklearn.ensemble*)

Es un modelo que combina múltiples árboles de decisión entrenados secuencialmente, donde cada árbol corrige los errores del anterior. Es eficaz para mejorar el rendimiento en problemas complejos y no lineales.

Clasificador	Tiempo de Entrenamiento	Accuracy Score	Observaciones
Regresión Logística	N/A	0.9867	Predijo correctamente la mayoría de las etiquetas, mostrando un desempeño muy alto en la clasificación de correos como spam o no spam.
Máquina de Soporte Vectorial (SVM)	55 segundos	0.795	Mostró un rendimiento inferior comparado con la regresión logística, sugiriendo que no separó óptimamente las clases en este caso.
Random Forest	2 segundos	0.9867	Tuvo un rendimiento similar a la regresión logística, mostrando ser un modelo eficaz para esta tarea.
Random Forest con Ajuste de Hiperparámetros	N/A	0.9867	El ajuste de hiperparámetros mejoró marginalmente el rendimiento, manteniendo el mismo accuracy. Los mejores parámetros fueron: `min_samples_leaf=1, min_samples_split=6, n_estimators=1000`
Gradient Boosting	226 segundos	0.9767	Tuvo un alto rendimiento con un AUC de 0.9986, lo que indica un excelente desempeño para separar las clases.

Cuadro 1
Resultados de los clasificadores

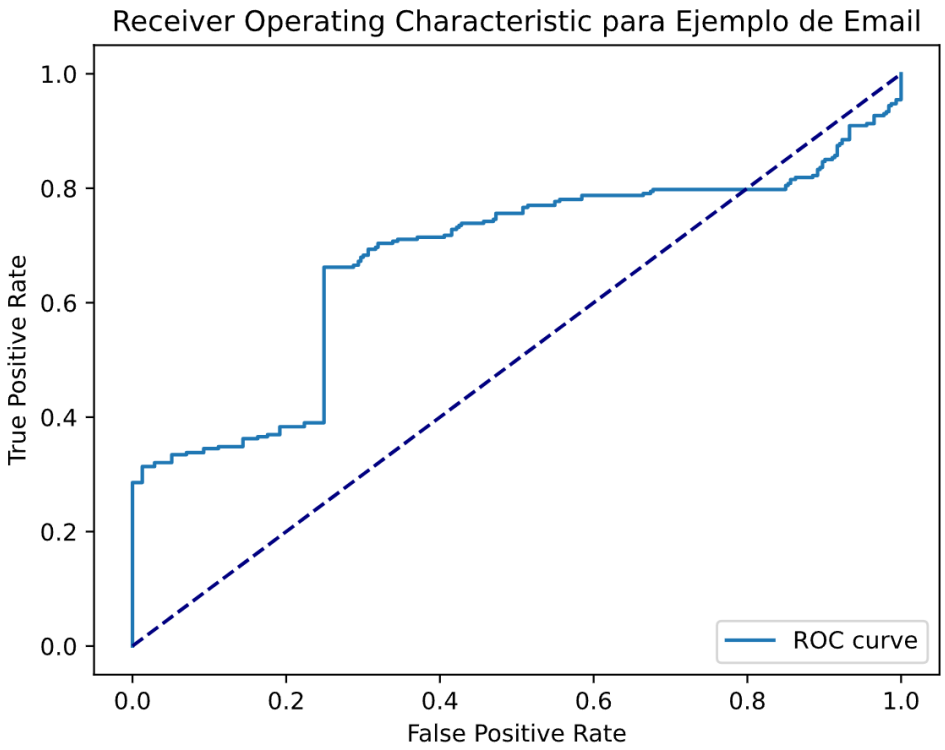

Figura 1
Curva ROC - Gradient Boosting

La curva ROC (Receiver Operating Characteristic) se utiliza para evaluar el rendimiento de un clasificador binario. En ella, el **eje X** representa la **Tasa de Falsos Positivos** (False Positive Rate) y el **eje Y** representa la **Tasa de Verdaderos Positivos** (True Positive Rate).

Curva Roc:

- La línea azul representa el rendimiento del modelo. Cuanto más cerca esté la curva del vértice superior izquierdo, mejor será el rendimiento del modelo, ya que esto significa que el modelo tiene una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.
- La línea discontinua en color azul marino que va de la esquina inferior izquierda a la superior derecha es la línea de "no discriminación.º aleatoriedad. Un modelo que sigue esta línea básicamente hace predicciones al azar.

En este caso, existe un desempeño aceptable, ya que la curva ROC se encuentra por encima de la línea de aleatoriedad. Además se obtuvo un AUC de 0.9986, indicando así un modelo casi perfecto. Sin embargo, la curva no llega completamente a la esquina superior izquierda, lo que sugiere que el modelo aún tiene margen para mejorar su capacidad de discriminar entre clases (spam y no spam).

Conclusión

El desempeño de los clasificadores muestra que, en general, los modelos implementados lograron una alta precisión en la clasificación de correos electrónicos como spam o no spam. La **regresión logística** y el **Random Forest** destacaron con un **accuracy del 98.67 %**, lo que indica una capacidad muy alta de predicción. El **Gradient Boosting** también tuvo un rendimiento sólido con un **accuracy del 97.67 %** y un excelente **AUC**, mostrando que es un modelo robusto para esta tarea.

Por otro lado, el **SVM** fue el que tuvo el peor desempeño relativo, con un **accuracy del 79.5 %**, lo que indica que no fue capaz de separar de manera efectiva las clases en comparación con los otros modelos. En general, los clasificadores tuvieron un rendimiento adecuado, con la regresión logística y Random Forest liderando los resultados.

Referencias

IBM. (2024, May 14). ¿Qué es la regresión logística? *IBM*. Retrieved from https://www.ibm.com/mx-es/topics/logistic-regression

Navlani, A. (2024, March 5). Tutorial sobre máquinas de vectores de soporte con Scikit-learn. Retrieved from https://www.datacamp.com/es/tutorial/svm-classification-scikit-learn-python

IBM. (n.d.). Random Forest. Retrieved from https://www.ibm.com/mx-es/topics/random-forest

Foqum Analytics. (2023, November 16). Gradient Boosting. *FOQUM*. Retrieved from https://foqum.io/blog/termino/gradient-boosting

Chan, C. (2024, February 22). What is a ROC Curve - How to Interpret ROC Curves. *Displayr*. Retrieved from https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/