Clasificación de email de spam: pre-procesamiento y baselines

Luis Maximiliano López Ramírez

September 2024

1 Comparación del desempeño de todos los clasificadores entrenados

Modelo	Accuracy	Tiempo de Entrenamiento (s)	Otras Métricas
Logistic Regression	0.9867	–	–
SVC	0.80	58	–
Random Forest	0.98	2	–
Random Forest (Optimizado)	0.9817	–	`min_samples_leaf: 1, min_samples_split: 6, n_estimators: 1000`
Gradient Boosting	0.965	240	`CV Score: Mean - 0.9941, Std - 0.0065`

Table 1: Comparación del Desempeño de los Modelos Clasificadores

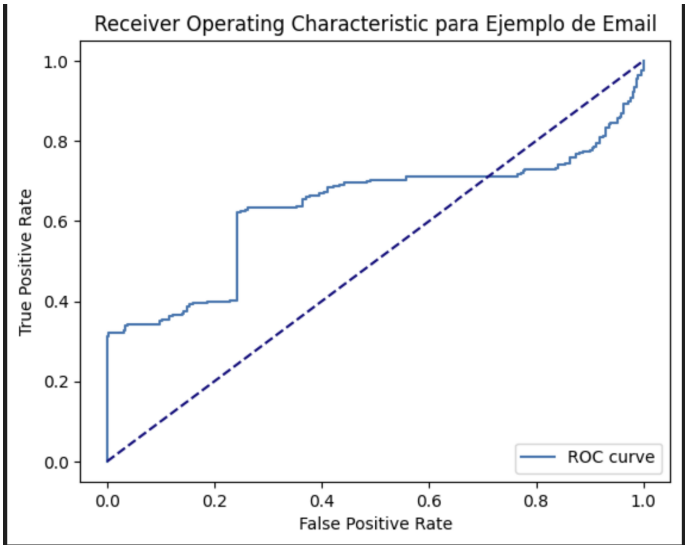

Figure 1: Enter Caption

Curva ROC (Receiver Operating Characteristic): se utiliza para evaluar el rendimiento de un modelo de clasificación binaria.

- **Eje X (False Positive Rate)**: Muestra la tasa de falsos positivos (clasificados incorrectamente como positivos). Va de 0 a 1.
- **Eje Y (True Positive Rate)**: Representa la tasa de verdaderos positivos o sensibilidad. También va de 0 a 1.
- **Línea azul (Curva ROC)**: Muestra cómo varían las predicciones al ajustar el umbral de clasificación.
- **Línea punteada (Base)**: Representa el rendimiento de un clasificador aleatorio (50-50).

Interpretación:

- Cuanto más cerca esté la curva del vértice superior izquierdo, mejor es el modelo.
- Una curva cercana a la diagonal indica un rendimiento cercano al azar.

Área bajo la curva (AUC):

- Cuantifica el rendimiento. Un AUC de 1 indica un modelo perfecto, mientras que un AUC de 0.5 indica un modelo aleatorio.

2 Interpretación de Resultados

La tabla anterior compara el desempeño de varios modelos de clasificación. A continuación se presentan las interpretaciones clave:

- **Logistic Regression:** Este modelo obtuvo la mejor exactitud (accuracy) con un valor de **0.9867**.
- **SVC:** Aunque tiene un buen desempeño con una exactitud de **0.80**, su tiempo de entrenamiento fue relativamente alto, con **58 segundos**.
- **Random Forest:** El modelo estándar de Random Forest tuvo una exactitud muy alta de **0.98**, con un tiempo de entrenamiento mucho más bajo (**2 segundos**) comparado con SVC. Este modelo parece ser más eficiente en cuanto a tiempo de procesamiento.
- **Random Forest (Optimizado):** Después de realizar una optimización de los hiperparámetros, el modelo Random Forest mejoró ligeramente su exactitud a **0.9817**. A pesar de que no se indica el tiempo de entrenamiento, las métricas adicionales muestran los hiperparámetros óptimos utilizados: `min_samples_leaf: 1`, `min_samples_split: 6`, `n_estimators: 1000`. Esto indica que la optimización mejoró el rendimiento, pero no significativamente en términos de exactitud.
- **Gradient Boosting:** Este modelo tiene una exactitud de **0.965**, menor que la de Random Forest y Logistic Regression. Sin embargo, el tiempo de entrenamiento fue considerablemente más alto, **240 segundos**. Esto sugiere que Gradient Boosting es más costoso computacionalmente. Además, la métrica de cross-validation muestra un rendimiento sólido con una puntuación de **CV Score: Mean - 0.9941, Std - 0.0065**, lo que indica que el modelo es bastante estable.

En general, **Logistic Regression** ofrece el mejor desempeño en términos de exactitud, mientras que **Random Forest** es una opción eficiente y competitiva en términos de tiempo de entrenamiento. **Gradient Boosting** parece ser adecuado cuando se prioriza la estabilidad, pero requiere mucho más tiempo de entrenamiento.