Sofía Cantú Talamantes A01571120

Machine Learning Classifiers on the Enron Email Dataset

Comparación de resultados

Resumen

Modelo	Precisión	Tiempo de formación
Regresión logística	98.67%	0 seconds
Support Vector Machine	79.67%	53 seconds
Random Forest	96.83%	1 second
Gradient Boosting	97.50%	93 seconds

Reporte

Modelos

Regresión logística
- Alta precisión con un coste computacional mínimo
- Excelente modelo de referencia para esta tarea de clasificación binaria
- El conjunto de datos parece adecuado para modelos lineales

Support Vector Machine
- Precisión significativamente inferior a la de otros modelos
- El mayor tiempo de entrenamiento sugiere ineficacia para este conjunto de datos
- Puede beneficiarse del ajuste de los hiper parámetros

Random Forest
- Alto rendimiento con un tiempo de entrenamiento muy rápido
- Resistente al sobreajuste
- Excelente equilibrio entre eficacia y precisión

Gradient Boosting
- Mayor precisión global y capacidad para captar patrones complejos
- Mayor tiempo de formación, pero mejores métricas de rendimiento
- La mejor opción cuando los recursos informáticos no son un problema

Sofía Cantú Talamantes A01571120

Conclusiones

El mejor modelo general es Gradient Boosting, considerando su precisión y capacidad para reconocer patrones. En cuanto a la eficiencia frente al rendimiento, si tomamos en cuenta la velocidad es el Random Forest, pero el Gradient Boosting tiene el máximo rendimiento. El rendimiento de Support Vector Machine dejó que desear, requiere una mayor investigación y posiblemente ajustes adicionales.

Recomendaciones

1. Ajuste de hiper parámetros para el modelo Support Vector Machine, en tiempo y rendimiento es bastante malo.
2. Ajuste de hiper parámetros para el modelo Random Forest para ver si puede ganarle al Gradient Boosting sin sacrificar la velocidad.