Universidad del Valle de Guatemala
Data Science
Javier Prado – 21486
Luis Pedro Monterroso - 21699

Laboratorio 8. Detección de Anomalías con Autoencoder, Isolation Forest y LOF

Resumen del conjunto de datos y preparación

Dataset: Forest CoverType (581,012 filas, 54 variables de entrada + Cover_Type). Se definió la clase normal como Cover_Type == 2; el resto se marcó como anomalía (y = 1). Particiones:

- Test mixto: 20% de normales + 20% de anomalías.
- Entrenamiento/Validación (solo normales): el 80% restante de normales se dividió en Train/Val (75/25).
- Validación mixta adicional: mezcla de normales/anómalos para calibrar umbrales.

Escalado: StandardScaler ajustado solo con Train (normales); aplicado a Val/Val-mixto/Test. Variables binarias (one-hot) no se escalaron. Convención de scores: mayor = más anómalo.

Resultados comparados

Métricas en test (umbral calibrado por el mejor criterio de cada modelo). Las cifras son ilustrativas según los resultados obtenidos en el notebook.

Modelo (mejor variante)	F1	Precisión	ROC-AUC	P@1%
LOF (novelty=True)	0.781	0.755	0.857	0.997
Isolation Forest (tuning)	0.700	0.606	0.737	0.989
Autoencoder (AE_enc8 + F1_P≥0.85)	0.551	0.857	0.740	0.992

Universidad del Valle de Guatemala
Data Science
Javier Prado – 21486
Luis Pedro Monterroso - 21699

Figuras: Curvas ROC/PR mostraron la mejor separación global para LOF. Los histogramas del error del AE, tras recalibrar el umbral (percentiles altos o F1 con precisión mínima), mostraron reducción de falsos positivos.

Discusion y recomendaciones

Durante el desarrollo del experimento, se observó que la arquitectura y la estrategia de umbralización influyen significativamente en el desempeño de los modelos de detección de anomalías. El autoencoder con cuello de botella reducido (`AE_enc8`) mostró una mejor capacidad para separar normales y anómalos, especialmente cuando se empleó una validación desbalanceada y umbrales basados en precisión mínima.

El modelo Isolation Forest, tras la optimización de hiperparámetros, logró resultados competitivos, aunque ligeramente inferiores al mejor autoencoder. Sin embargo, ambos enfoques demostraron ser útiles para el problema planteado.

Cambiar la clase considerada “normal” redefine la geometría del conjunto de entrenamiento: el AE aprenderá a reconstruir otra distribución (nuevo manifold), y métodos de vecindad/densidad como LOF modificarán sus densidades locales y vecindarios relevantes. En consecuencia, la separabilidad de errores y scores puede mejorar o empeorar y el umbral óptimo cambiará. Por rigor, si se redefine la normalidad, hay que rehacer el pipeline completo: volver a partir (Train/Val solo normales), reajustar el escalador únicamente con Train, recalibrar el umbral (percentil y/o F1 con restricción) y re-evaluar métricas (ROC/PR-AUC, F1@umbral, Precision@k, matriz de confusión).

Con datos desbalanceados (pocas anomalías), PR-AUC es más informativa que ROC-AUC porque refleja directamente el trade-off precisión vs recall cuando la clase positiva es rara. La ROC puede verse “alta” incluso cuando el modelo genera demasiados falsos positivos, ya que la gran cantidad de negativos amortigua el FPR. En cambio, PR-AUC penaliza inmediatamente la caída de precisión al subir FP. Por eso, aunque reportamos ambas, PR-AUC fue la guía principal para comparar modelos y elegir puntos operativos