BreakNbuild - Una herramienta para evaluar el efecto del tamaño de la muestra en modelos de Machine Learning

Palabras clave: Machine Learning, Curva de aprendizaje, Evaluación de modelos

Autor: Francisco Cardozo, Msc.

Abstract

El rendimiento de los modelos de machine learning depende de varios factores, como la calidad de los datos, la complejidad del modelo, las características de los algoritmos y la cantidad de datos disponibles para el entrenamiento. Es común que los datos suelan ser limitados o están en constante cambio, por lo que es esencial evaluar cómo las variaciones en el tamaño de la muestra afectan la precisión y estabilidad del modelo. Sin esta comprensión, los modelos corren el riesgo de ser menos efectivos o poco fiables, especialmente cuando se utilizan para predecir condiciones que cambian constantemente.

Para enfrentar este desafío, BreakNBuild es un paquete de R diseñado para analizar cómo el rendimiento de los modelos de machine learning responde a cambios progresivos en el tamaño de los datos de entrenamiento (learning curves). La función principal del paquete, *progressive_splits,* permite simular escenarios donde la cantidad de datos disponibles varía gradualmente, facilitando la creación de experimentos para evaluar la sensibilidad del modelo ante diferentes tamaños de muestra. Además, BreakNBuild ofrece herramientas de visualización que permiten explorar el impacto del tamaño de la muestra en diversas métricas de rendimiento estándar en machine learning, integrándose perfectamente en flujos de trabajo basados en el ecosistema Tidymodels.

Durante la presentación, se ilustrará el uso de BreakNBuild mediante tres casos de estudio con datos simulados. En el primer caso, se mostrará cómo funciona el paquete con una base de datos generada aleatoriamente. En el segundo, se trabajará con un conjunto de datos donde las variables tienen una relación lineal, y en el tercero, con una relación no lineal entre variables. Estos ejemplos destacarán cómo diferentes algoritmos responden a variaciones en el tamaño de la muestra y cómo los errores de los modelos cambian a medida que se incrementa el número de observaciones, brindando una comprensión más profunda del comportamiento del modelo bajo distintos escenarios.

Subtítulo
Análisis de Curvas de Aprendizaje con BreakNBuild

Subtítulo 2
Mejora de Modelos Basada en Curvas de Aprendizaje