Actividad 2 – Machine Learning II

Profesor: Francisco Pérez Galarce
Formato de entrega. Un notebook en Python con: Código limpio, celdas ordenadas y comentarios. Gráficos y tablas de resultados.
Dentro del mismo notebook usando Markdown debe incluir: Descripción del preprocesamiento. Comparación de modelos. Discusión de resultados y conclusiones.
La actividad puede ser desarrollada en grupos de máximo 3 personas.
El avance de cada grupo será discutido en la próxima clase, martes 16 de diciembre.

CONTEXTO

En la actividad anterior se abordó el problema de churn utilizando modelos lineales (regresión logística), analizando el impacto del preprocesamiento, la regularización y las transformaciones polinomiales.
En esta actividad se busca extender el análisis hacia modelos no lineales basados en árboles, con énfasis en:

- Selección sistemática de hiperparámetros mediante Grid Search Cross-Validation y Random Search Cross Validation.
- Comparación de desempeño predictivo y costos computacionales de los métodos para seleccionar hiperparámetros.
- Análisis de varianza de las predicciones en modelos de ensamble (Random Forest).
- Evaluación mediante métricas de clasificación, curvas ROC y Precision–Recall.

OBJETIVOS DE APRENDIZAJE

Al finalizar la actividad, la o el estudiante será capaz de:

- Ajustar y optimizar un árbol de decisión de clasificación usando validación cruzada.
- Interpretar y visualizar el árbol de decisión seleccionado.
- Analizar el efecto del número de árboles en la varianza de un Random Forest.
- Seleccionar un Random Forest mediante exploración de hiperparámetros.
- Comparar modelos no lineales usando métricas robustas para problemas desbalanceados.

INSTRUCCIONES

Paso 1. Árbol de decisión

- Implementar un árbol de decisión de clasificación (DecisionTreeClassifier) utilizando el mismo conjunto de datos y preprocesamiento definidos en la Actividad 1.
- Definir una grilla de hiperparámetros, al menos: max_depth, min_samples_split, min_samples_leaf, criterion (gini, entropy o log_loss). Justifique los rangos o valores posibles para la selección de los valores de las grillas, considerando aspectos de sobreajustes.
- Utilizar Grid Search Cross Validation y Random Search Cross Validation para: (i) seleccionar el árbol de clasificación óptimo, (ii) usar como métrica principal F1 o PR-AUC de la clase churn.
- Compare los tiempos de los dos métodos de búsqueda de hiperparámetros y las métricas de clasificación.
- Seleccione los parámetros para el mejor árbol de decisión para clasificación.

Paso 2. Visualización e interpretación del árbol óptimo

- Visualizar el árbol de decisión seleccionado: Usar plot_tree. Limitar la profundidad visualizada si es necesario para legibilidad.
- Variables más relevantes en las primeras divisiones. Compare la interpretabilidad del modelo frente a la regresión logística.

Paso 3. Random Forest y análisis de varianza

- Implementar validación cruzada k-fold estratificada para Random Forest.
- Estudiar explícitamente la varianza de las predicciones al variar el número de árboles. Entrenar modelos con: 2, 4, 8, 16, 32, 64 y 128 árboles. Para cada configuración:
    - Calcular la varianza de las probabilidades predichas entre folds.
    - Registrar métricas de desempeño (F1, AUC-ROC, PR-AUC).
- Graficar varianza de las predicciones vs. número de árboles.
- Graficar métricas de clasificación vs. número de árboles.
- Discutir relación entre número de árboles, estabilidad y costo computacional. Evidencia empírica de reducción de varianza.

Paso 4. Selección del mejor Random Forest

- Definir una grilla de hiperparámetros para Random Forest, por ejemplo: n_estimators, max_depth, min_samples_leaf, max_features.
- Usar Grid Search Cross-Validation o Random Search Cross Validation para seleccionar el mejor modelo..
- Comparar los modelos seleccionados Árbol de decisión vs. Random Forest.

Paso 5. Comparación final y análisis crítico

- Trate de mejorar los modelos seleccionados mediante el uso de pesos por clase.
- Para el mejor árbol de decisión y el mejor Random Forest reportar métricas de clasificación: Accuracy, Precision, Recall, F1-score.
- Analizar y graficar curva ROC y AUC-ROC.
- Analizar y graficar curva Precision–Recall y PR-AUC.
- ¿Qué modelo es más adecuado para el problema de churn?
- ¿Random Forest supera al árbol individual?¿Por qué?
- Impacto del desbalance de clases en la interpretación de resultados.

Paso 6. Análisis crítico

- Discuta la relación entre varianza, ensambles y generalización.
- ¿Qué ventajas y desventajas presentan los árboles frente a modelos lineales en este problema?
- ¿En qué casos preferiría un árbol interpretable sobre un Random Forest?
- ¿Cómo se relaciona la reducción de varianza observada con la teoría vista en clases?
- Desde una perspectiva de negocio, ¿qué métrica priorizaría para campañas de retención y por qué?