Instituto Tecnológico y de Estudios Superiores de Monterrey

TC3006C.103

Inteligencia Artificial Avanzada para la Ciencia de Datos I

Ingeniería en Ciencias de Datos y Matemáticas

Momento de Retroalimentación: Módulo 2 - Análisis sobre el Desempeño del Modelo

Profesor
Jesús Adrián Rodríguez Rocha

Alumno

Alejandro José Murcia Alfaro
A00828513

Monterrey, Nuevo León. 10 de septiembre de 2023

1. Análisis y Optimización del Modelo de Clasificación

1.1. Separación y Evaluación del Modelo

Para garantizar una evaluación precisa y prevenir el sobreajuste, se dividió el conjunto de datos en tres segmentos distintos: entrenamiento, validación y prueba. Esta división es esencial para evaluar la capacidad del modelo para generalizar a datos no vistos. A continuación, se detalla el rendimiento de cada modelo en estos conjuntos:

Modelo Inicial

El modelo inicial se entrenó con parámetros predeterminados para establecer una línea base de rendimiento. Los resultados obtenidos fueron:

- **Accuracy en entrenamiento:** 78.26 %
- **Accuracy en validación:** 74.68 %
- **Accuracy en prueba:** 77.27 %

Este modelo proporciona una comprensión inicial del comportamiento del algoritmo sin ningún tipo de optimización. Aunque los resultados son aceptables, es evidente que hay margen para mejorar, especialmente en términos de generalización a datos no vistos.

Modelo Optimizado

Para el modelo optimizado, se realizó una búsqueda exhaustiva de hiperparámetros utilizando 'GridSearchCV'. Esta técnica permite encontrar la mejor combinación de parámetros para maximizar el rendimiento del modelo. Los resultados obtenidos fueron:

- **Accuracy en entrenamiento:** 99.57 %
- **Accuracy en validación:** 77.92 %
- **Accuracy en prueba:** 76.62 %

A pesar de su impresionante rendimiento en el conjunto de entrenamiento, el modelo optimizado no logró generalizar adecuadamente a los conjuntos de validación y prueba. Esta discrepancia indica un claro sobreajuste, donde el modelo se ha adaptado demasiado a los datos de entrenamiento.

Modelo Ajustado

Dado el sobreajuste observado en el modelo optimizado, se decidió ajustar y regularizar el modelo. Se ajustaron parámetros como maxdepth, minsamplessplit, y minsamplesleaf basándose en el análisis previo y en las gráficas de Bias-Variance Trade-off. Los resultados obtenidos fueron:

- **Accuracy en entrenamiento:** 80.43 %
- **Accuracy en validación:** 77.27 %
- **Accuracy en prueba:** 77.27 %

El modelo ajustado logró un equilibrio entre el sesgo y la varianza, mostrando un rendimiento consistente en todos los conjuntos de datos.

1.2. Análisis Gráfico de Bias-Variance Trade-off

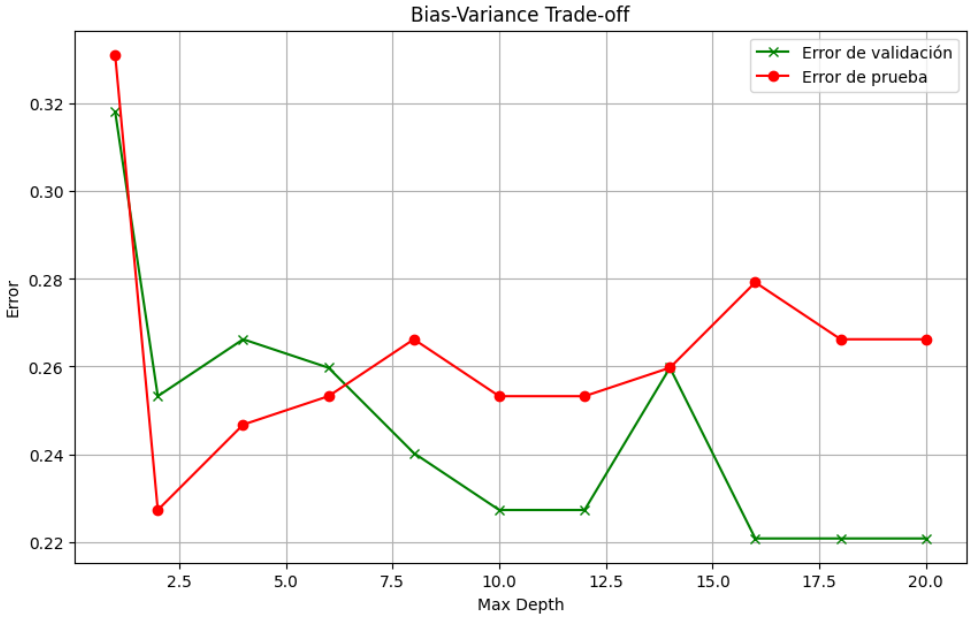

Figura 1: Bias-Variance Trade-off Modelo Inicial y Optimizado

El gráfico anterior muestra el error de validación y prueba para el modelo inicial y optimizado en función de la complejidad del modelo (representada por 'max depth'). Al inicio, ambos errores disminuyen, alcanzando un mínimo en un 'max depth' de aproximadamente 2.0. Esto sugiere que hasta este punto, el modelo mejora su capacidad de generalización.

Después de este punto, el error de validación comienza a aumentar, mientras que el error de prueba se mantiene más o menos constante. Esta divergencia indica un aumento en la varianza del modelo, lo que sugiere sobreajuste. El modelo está capturando demasiado ruido o variabilidad en el conjunto de entrenamiento, lo que afecta su capacidad para generalizar a nuevos datos.

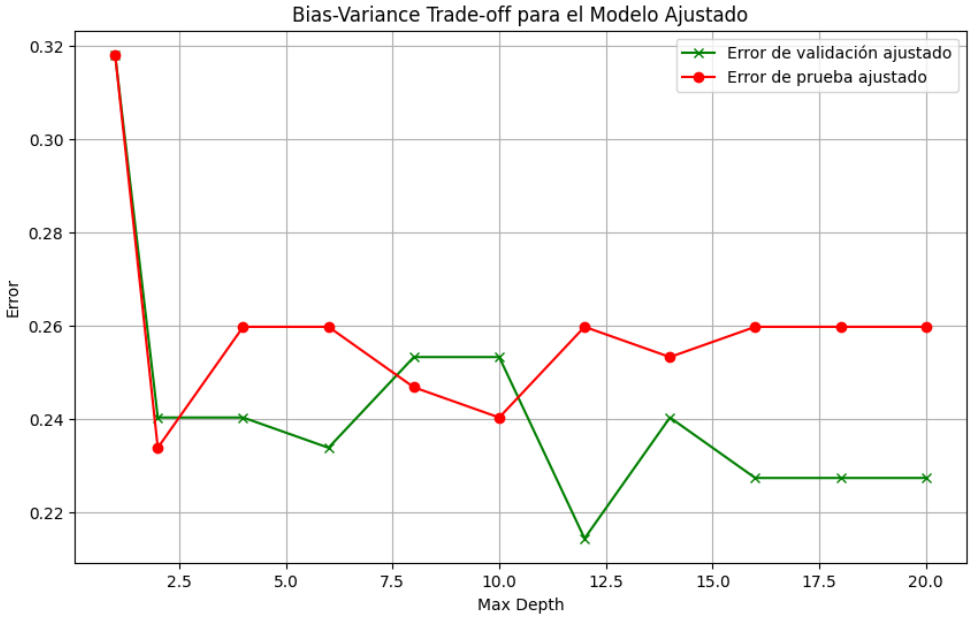

Figura 2: Bias-Variance Trade-off Modelo Ajustado

El segundo gráfico muestra el comportamiento del modelo ajustado. A diferencia del modelo inicial y optimizado, el modelo ajustado muestra una tendencia más estable después del punto de ‘max depth‘ de 2.0. Aunque hay un ligero aumento en el error de validación, es menos pronunciado que en el modelo optimizado. Esto sugiere que los ajustes realizados al modelo han reducido su varianza, haciendo que sea más robusto y generalice mejor.

Basándonos en estos gráficos, podemos inferir que el modelo inicial y optimizado tiene un sesgo medio y una varianza alta, mientras que el modelo ajustado tiene un sesgo medio y una varianza media.

1.3. Conclusión

A través de este análisis, hemos evaluado y optimizado un modelo de clasificación basado en el algoritmo Random Forest. El modelo inicial mostró un rendimiento decente, pero había margen de mejora. El modelo optimizado, aunque mostró un rendimiento casi perfecto en el conjunto de entrenamiento, no logró generalizar bien en los conjuntos de validación y prueba. Finalmente, el modelo ajustado, después de aplicar técnicas de regularización y ajuste de parámetros, logró un buen equilibrio entre sesgo y varianza, mostrando un rendimiento consistente en todos los conjuntos de datos.

Lo ideal en este tipo de problemas es utilizar el modelo ajustado para utilizarlo al hacer predicciones o utilizar el modelo, ya que logra un buen equilibrio entre sesgo y varianza y muestra un rendimiento consistente en todos los conjuntos de datos.