Desempeño del modelo

Alex Arguelles Elias

A00829536

Análisis del modelo

Introducción:

En este reporte analizamos los pasos que seguimos para entrenar nuestro modelo de machine learning paso por paso y como terminamos creando un modelo de la mejor manera posible, utilizando técnicas para poder combatir una alta varianza o un bajo sesgo, también como llegamos a los hiperparametros mas adecuados para el modelo

Definición del Problema:

Estamos trabajando con un problema de regresión, por lo tanto usaremos un árbol de regresión

Nuestros datos consisten en 1600 ejemplos, cada uno con 12 características.

Usaremos el error cuadrático medio como nuestra métrica de evaluación.

Preprocesamiento de Datos:

Limpieza de datos: Eliminamos los valores nulos y no normalizamos los datos debido al modelo que estamos usando que no es sensible a eso.

División de datos: 80% entrenamiento, 20% prueba.

Selección del Modelo:

Escogí el árbol de regresión debido a que ya eh usado el de clasificación y quería revisar si era diferente su uso y ver que tal me iba con él, además que no es sensible a la normalización y todo lo que sea ahorrarme limpieza de datos es un plus

Optimización de Hiperparámetros:

2 maneras de encontrar hiperparámetros:

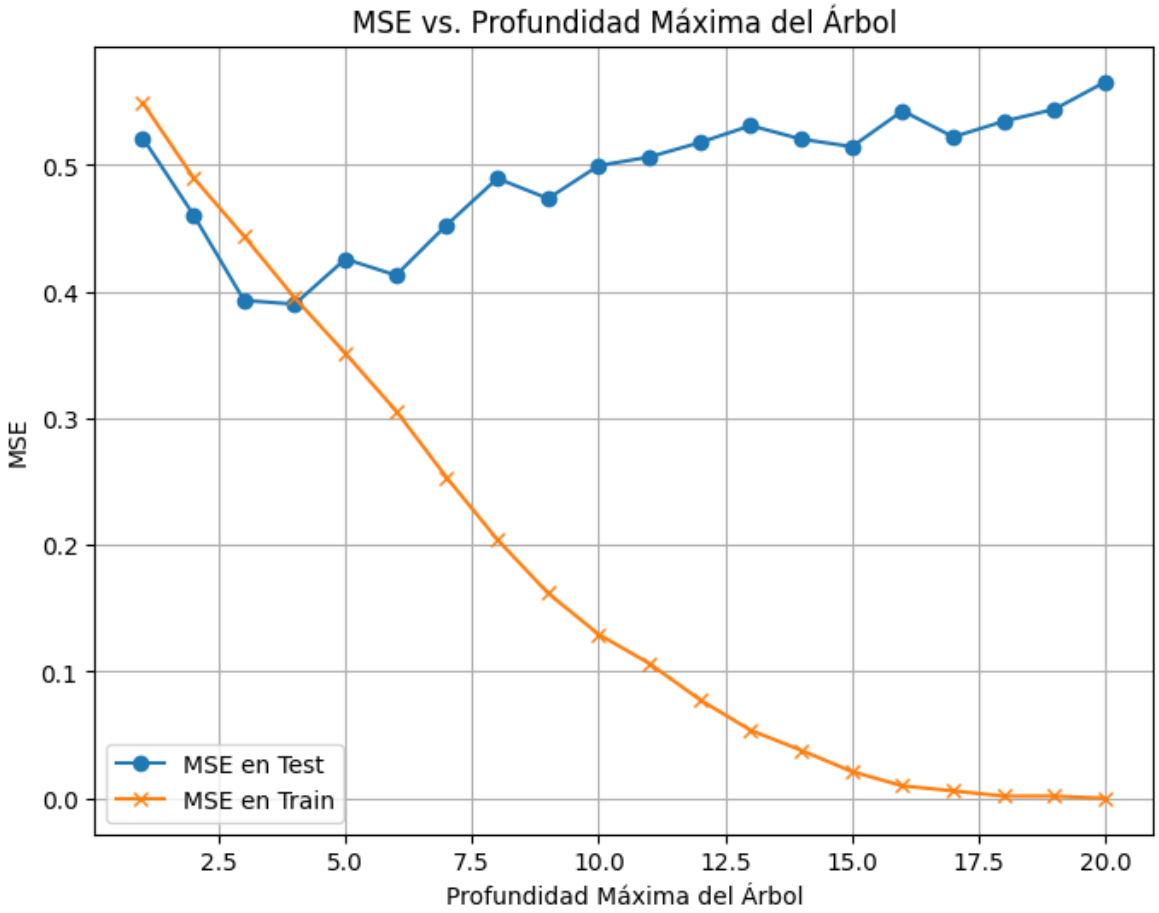
- Utilizaremos la vieja estrategia de probar diferentes combinaciones en un inicio para verificar que todo este funcionando como corresponde y con este modelo graficaremos la tabla que mas adelante nos servirá para analizar el entrenamiento del modelo
- Usaremos Grid Search para probar diferentes combinaciones de hiperparámetros. Por ejemplo, probaremos diferentes valores para la profundidad máxima, el criterio a utilizar y las hojas mínimas

Entrenamiento y Validación:

Entrenamos el modelo con diferentes combinaciones de hiperparámetros y validamos su rendimiento en el conjunto de validación.

Análisis de Varianza y Sesgo:

Para analizar nuestros valores usaremos la siguiente tabla:



Diagnóstico y explicación el nivel de ajuste del modelo: underfitt fitt overfitt

Podemos ver en la gráfica como se desempeña el modelo en los datos de train y test (menor error es mejor) curiosamente el modelo de prueba da un mejor rendimiento al principio, pero si no entrenamos el modelo hasta que tenga una profundidad de 4 no alcanzamos a reducir de mejor

manera el error, si solo usamos una profundidad de 3 o menor el modelo sigue teniendo un nivel alto de error lo que nos indicaría que tiene underfitting o le falta entrenamiento

pero si seguimos entrenando el modelo podemos ver como reducimos el error de set de prueba hasta 0 pero el error de el set prueba aumenta, lo cual nos indica que nuestro modelo se esta sobre entrenando y no funcionada en el mundo real, solo funciona con los datos que ya conoce

Diagnóstico y explicación el grado de varianza: bajo medio alto

cabe recalcar que sabemos que la grafica tambien nos indica la varianza, al principio la varianza es baja por que el mode esta en underfitting, pero conforme aumenta la complejidad observamos que aumenta la varianza entre el set de entrenamiento y de prueba

Diagnóstico y explicación el grado de bias o sesgo: bajo medio alto

y por último el Bias aumenta conforme el modelo entiende mejor el data set, por eso al principio es bajo e ira aumentando, el problema principal es que el modelo se sobre entrenara lo cual hace el error de bias baje pero aumentando la varianza

Resumen

podemos revisar que se cumple lo visto en clase, al principio la varianza es bajar y el sesgo es alto mientras el modelo está en underfitting pero conforme avanza revisamos que la varianza aumenta, el sesgo disminuye hasta que están en equilibrio cuando esta el modelo correctamente entrenado y si lo sobre entrenamos vemos que la varianza se dispara y el sesgo se reduce

Regularización:

Después de analizar nuestro modelo vemos cuales son los hiperparametros que podemos utilizar y escojemos un punto donde nuestro modelo este correctamente entrenado (con una profundidad de 4 en el primero y 5 usando grid search y otro hiperparametros)

Conclusión:

Con las técnicas utilizadas reducimos el error de una manera significativa, cuidando tanto el fit del modelo como la varianza y el sesgo, dando un resultado deseable