Análisis y Reporte sobre el Desempeño del Modelo

Russel Rosique Rodríguez
A01283727

En esta entrega intermedia, se utilizó el dataset de kaggle llamado “Students performance in exams”. Este dataset consiste de 8 columnas: “gender”, “race/ethnicity group”, “parental level of education”, “lunch”, “test preparation course”, “math score”, “writing score”, “reading score”. Se decidió implementar un modelo de aprendizaje para este dataset, de tal forma que el modelo pudiera predecir el promedio de los tres puntajes de un alumno, en base a el resto de las columnas. Se consideró a este conjunto de datos como apto para la implementación ya que consiste de un total de 1000 datos únicos, y registrados de forma completa.

Separación y Evaluación del Modelo

Como modelo, se eligió utilizar la herramienta de random forests, ya que es un gran modelo de aprendizaje, además que cuenta con diversas ventajas. Para evaluar este modelo con el conjunto de datos, se realizó un split de los datos en un conjunto de entrenamiento y uno de evaluación. Se tomaron el 80% de los datos de forma aleatoria para el conjunto de entrenamiento, y el 20% restante para el conjunto de evaluación. Una vez hecho esto, se implementó el modelo con el algoritmo de random forest con los siguientes parámetros:

$$nestimators = 15$$
$$maxdepth = 9$$
$$randomstate = 42$$

Posterior a esto se realizó la evaluación del modelo con ambos conjuntos de datos (entrenamiento y prueba). Con el conjunto de entrenamiento, se obtuvieron las siguientes métricas:

MSE= 121.92537893770688
MAE= 8.909239624927874
R2= -0.7067617240634501

Por otro lado, con el conjunto de prueba se obtuvieron las siguientes métricas:

MSE= 215.06806740794644
MAE= 11.36903290962244
R2= -1.9521491364554615

Diagnóstico de Sesgo

Una vez obtenidos los resultados de la predicción del modelo, se comienza a diagnosticar el grado de sesgo que tiene el modelo. Para modelos de Random Forest, es común que se tenga un bajo grado de sesgo, mientras que la varianza suele ser elevada. No obstante, se realizó un rápido cálculo estimado del sesgo del modelo, comparando la media de la predicción del modelo, contra cada dato de la muestra tanto de entrenamiento como de prueba, mejor conocido como MAE.

En la comparación con la muestra de entrenamiento, se obtuvo este resultado: *8.9*
Por otro lado, con la muestra de test, se obtuvo: *11.36*

Dado que el rango de las predicciones se debe de encontrar entre 0 y 100 al ser calificaciones de exámenes, se analiza que el grado de sesgo es medio. Es un poco elevado para un modelo de random forest, sin embargo hay ciertas cosas que se pueden hacer para reducir el grado de sesgo.

Tras un análisis rápido, algo que evidentemente está elevando el sesgo es que se está incluyendo la columna de "raza" en el modelo de aprendizaje, cuando en realidad esta columna no es relevante para el modelo.

Diagnóstico de Varianza

Por otra parte, en los modelos de random forest la varianza suele ser alta, y suele causar problemas de overfitting. Para definir el grado de varianza de este modelo, se compararon los resultados entre los conjuntos de datos de train y test. Al comparar los MSE, se logra notar que existe una amplia diferencia entre ellos, lo que indica que se tiene un grado de varianza alto para el modelo.

Como se mencionó anteriormente, los modelos de random forest suelen tener un alto grado de varianza, por lo que se analiza que existe un área de mejora en este aspecto. Es posible realizar ciertos ajustes a los hiper parámetros del modelo para obtener mejores resultados y reducir esta varianza entre las evaluaciones.

Diagnóstico de Ajuste

Como ya se mencionó en el análisis de la varianza, se obtuvo un notorio mejor resultado al evaluar el modelo con el conjunto de entrenamiento, que al evaluarlo con el conjunto de prueba. Dicho esto, se tiene un poco de "overfitting" debido al alto grado de varianza. Se realizarán cambios al modelo para mejorar todo lo mencionado anteriormente, y se reportarán los resultados a continuación.

Cambios al Modelo

Tras el análisis y diagnóstico del grado de sesgo, varianza y ajuste del modelo, fue evidente que se tenían que hacer varios ajustes para encontrar un mejor balance entre el sesgo y la varianza, y terminar con un modelo con un ajuste correcto, ya que se tenía un notorio sobreajuste. Los cambios que se realizaron fueron los siguientes:

- Se eliminó la columna de "raza" por falta de relevancia, para disminuir el sesgo del modelo.
- Se cambió la columna de "average" que consistía del promedio entre los puntajes de matemáticas, escritura, y lectura. Se cambió a que consistiera únicamente del promedio entre lectura y escritura.
- Se cambió el tamaño de las muestras de entrenamiento y prueba. Anteriormente estaba dividido en 80:20, y cambió a 85:15. Esto con el propósito de disminuir la varianza.
- Se modificó el número de estimadores y la "max_depth" del modelo de random forest. Se analizó que se tenían muy pocos estimadores, lo que podía estar causando sobreajuste.

Resultados

Finalmente, se ejecutó el modelo nuevamente y generó nuevas predicciones con los cambios hechos. Se obtuvieron resultados positivos con los cambios, por lo que se denota que los cambios propuestos fueron pertinentes para mejorar el desempeño del modelo.

Con estos nuevos hiper parámetros:

$$nestimators = 50$$
$$maxdepth = 5$$
$$randomstate = 52$$

Se vieron los siguientes cambios para el conjunto train:

MSE= 121.92537893770688
MAE= 8.909239624927874
R2= -0.7067617240634501

Cambiaron a

MSE= 149.67566129396909
MAE= 9.772176068937158

R2= -1.647214021255555

Y se vieron los siguientes cambios para el conjunto de test:

MSE= 215.06806740794644
MAE= 11.36903290962244
R2= -1.9521491364554615

Cambiaron a

MSE= 185.79124508397038
MAE= 11.123896634489041
R2= -1.9542750091632892

Al analizar lo anterior, es evidente que aunque el sesgo en el modelo no disminuyó, se logró disminuir la varianza, logrando de esta manera alcanzar un mejor “fit” para los datos. Con el nuevo modelo, la diferencia entre el MSE de los distintos conjuntos es mínima, lo que indica un buen grado de ajuste.

No obstante, se analiza que el MSE en ambos conjuntos se mantiene alto, dado a que no existió un previo análisis exploratorio y extensa limpieza de los datos. Para obtener mejores métricas, y mejor precisión en las predicciones, sería necesario realizar este proceso, así como utilizar distintos modelos de aprendizaje y compara sus distintas ejecuciones.