TECNOLÓGICO DE MONTERREY

TC3006C: Inteligencia artificial avanzada para la ciencia de datos

Momento de Retroalimentación: Reto Análisis y Reporte sobre el desempeño del modelo

Francisco José Joven Sánchez - A00830564

Ingeniería en Tecnologías Computacionales

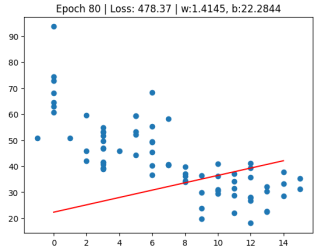
Fecha de entrega - 10 / 09 / 2023os

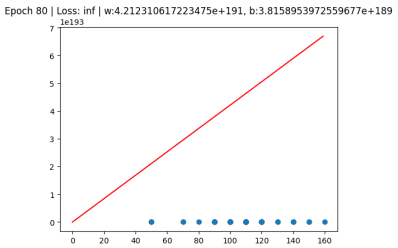
En la primera entrega de retroalimentación realice una regresión lineal simple con la cual predecir los valores de calificación que recibían los cereales en base a sus otras características.
En la segunda entrega de retroalimentación realice la implementación de un modelo de random forest para la predicción de cantidad de vistas de varios youtubers en el año 2023, el dataset consta de 1000 datos compuestos de muchos features.

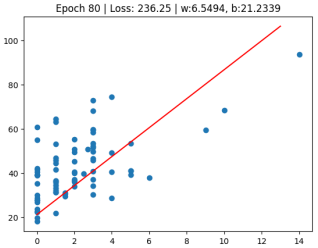
Primero realicé la limpieza en los datos para ambas entregas, quitando features, llenando nulos, creando dummies, pero para más detalles de ese apartado está el código en github, ya que lo importante de este reporte está en cuanto al modelo.

Lo primero que hay que hacer para el modelo es tener en cuenta su tipo, dado que el modelo de la primera entrega es sería una regresión lineal por lo que se entrena matemáticamente durante épocas que se le programe, en cuanto al segundo que utiliza random forest, este requiere entrenamiento a diferencia de algoritmos como el knn que también se utilizan para clasificar, requiere entrenamiento, para esto utilicé la función ya definida dentro de la librería de sklearn, la cual divide un dataset en 2 partes, una muestra del entrenamiento y otra de prueba, dado que la función tiene un factor de aleatoriedad, cada vez que se corre el programa se obtiene una división diferente de los datos. Ahora la validación se maneja desde el grid de parámetros, otra función de sklearn, en la que viene la opción de hacer cross validation.

En cuanto a los demás apartados necesarios para el reporte, es más adecuado utilizar la primera entrega al poder ser expresadas mejor con un modelo lineal que un algoritmo como el random forest.
En el primer entregable me enfoque en 3 features, en el primero se puede observar tanto una varianza como un bias medios, esto ya que el modelo requiere inclinarse hacia el lado opuesto y la altura para que se corrigiera en el punto en el que se acoplara mejor es más arriba, sin embargo no se siente como un gran salto a realizar para que sea bajo. El segundo tiene el mismo problema pero a una escala totalmente diferente y para mal, en este caso con cada iteración el gráfico se desvía inmensamente aumentando su error en proporciones gigantescas, por lo que se puede apreciar fácilmente que la varianza y el bias es muy alto. Finalmente el tercero si bien tiene un bias y varianza que se le puede considerar medio, varios epochs más y podría llegar a bajo, aunque no muy bajo, ya que se dispersa un poco más al principio que al final.

1.- 

2.- 

3.- 

Ahora en cuanto al fitting de cada una se puede decir que los modelos de la primera entrega están en todos sus apartados, el primero se encuentra en underfitting, dado que todavía puede realizar una mejora considerable en cuanto a su error, lo mismo con la tercera, aunque en su caso es más notable ya que podría estar muy cercas de ser fitting, para la segunda también debería ser underfitting, pero con la regresión lineal simple empeoraba en cada época, por lo que por este método se le podría considerar overfitting, pero es más que nada porque requiere un cambio de método. En cuanto al random forest no sabría ubicar en qué estaría, ya que por un lado están las las muestras de datos que como mencioné anteriormente cambian al correr el programa, pero en general me han estado dando principalmente en el rango de entre 15% - 50%, por lo que creo que se podría dar mejoría