Tecnológico de Monterrey

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Diego Armando Ulibarri Hernandez A01636875 ITC

Fecha: 13/09/22

Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para el análisis de datos

Análisis final del modelo regresión lineal

Para el dataset utilizado decidí utilizar la regresión lineal dado que considero que estos se prestan y se pueden adecuar correctamente a dicho modelo puesto que son datos numéricos y queremos predecir qué variables son las más adecuadas para utilizar en el modelo.

Los modelos lineales múltiples siguen la siguiente ecuación:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}) + e_i$$

Donde:

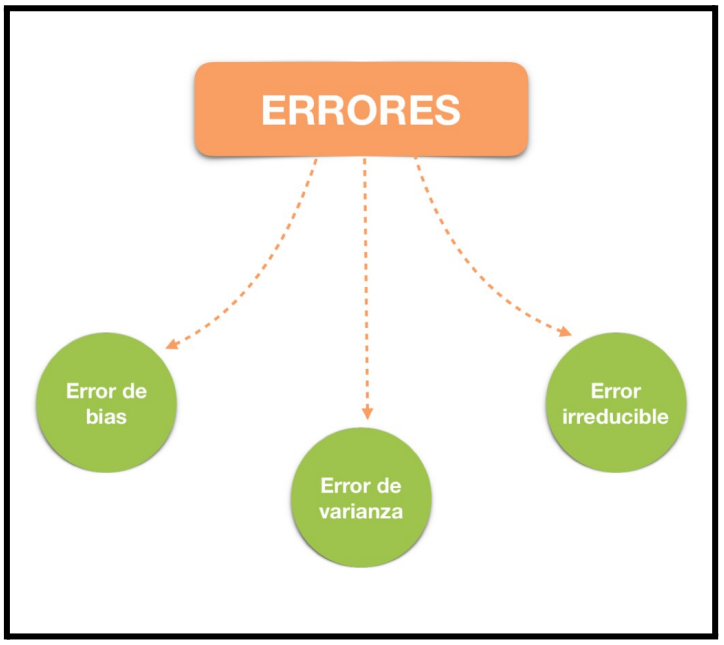
- β0: es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.
- βi: es el efecto promedio que tiene el incremento en una unidad de la variable predictora Xi sobre la variable dependiente Y, manteniéndose constantes el resto de variables. Se conocen como coeficientes parciales de regresión.
- ei: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

La regresión lineal múltiple trata de ajustar modelos lineales o linealizables entre una variable dependiente y más de una variables independientes.
Para obtener las variables independientes (x) tome a partir de mi propio criterio dos columnas relacionadas con la variable independiente (y), para este ejercicio no se realizó una limpieza de datos profunda, sin embargo el dataset no tiene valores NAN por lo que el trato con el mismo fue más sencillo, cabe recalcar que de haber dado un tratamiento adecuado al dataset los valores obtenidos en R^2 habrían sido mayores.
Se dividió el dataset en 4 variables:

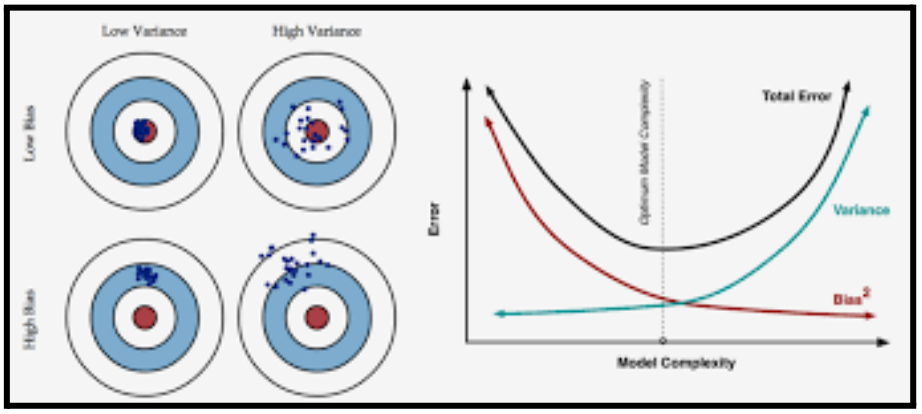
1. X_train: Contiene los valores de las variables independientes del 80% del dataset
2. X_test: Contiene los valores de la variable dependiente del 80% del dataset
3. y_train: Contiene los valores de las variables independientes del 20% del dataset
4. y_test: Contiene los valores de la variable dependiente del 20% del dataset

Esto ya que se utilizaron las dos primeras variables para la etapa de entrenamiento del modelo y las dos últimas para probar que tan bueno era el modelo haciendo predicciones.

En un modelo existen 3 tipos de errores:



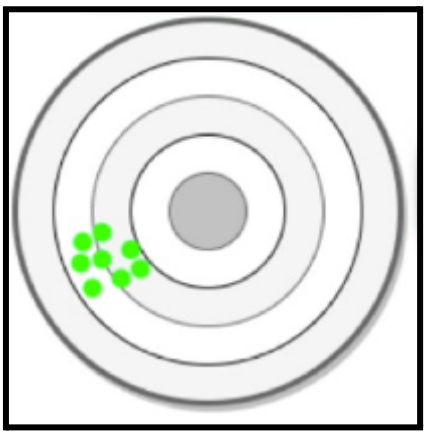
Cada error tiene un “nivel” que se puede englobar en alto o bajo, dependiendo de dicho nivel podemos observar tendencia de los datos de esta manera:



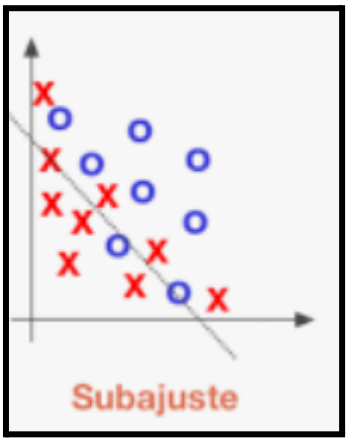
Los resultados obtenidos en mi modelo para las variables de sesgo y varianza son las siguientes:

```
MSE de bias_variance lib [pérdida promed
io esperada]: 55.433
/------------------------------/
Sesgo promedio: 53.651
/------------------------------/
Varianza promedio: 1.782
/------------------------------/
/------------------------------/
/------------------------------/
```

Como podemos observar, nuestra varianza es de 1.78 lo cual indica un valor bajo, por lo que podríamos decir que nuestra varianza es baja mientras que nuestro sesgo promedio es 53.65, representando un valor alto, esto entraría en la categorización vista en la imagen anterior



Y representa un sub ajuste o un underfitting:



Para mejorar el modelo, como antes mencionaba, se deben de tratar los datos de tal manera que puedan ajustarse bien al modelo y permitir una buena predicción, por ejemplo, escalarlos o normalizarlos, así como variar hiperparametros, tales como el número máximo de iteraciones o épocas.