Tecnológico de Monterrey

Análisis y Reporte sobre el desempeño del modelo.

Eduardo Rodríguez López A01749381

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Profesor: Jorge Adolfo Ramírez Uresti

13 de septiembre de 2022

Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Estado de México
Escuela de Ingeniería y Ciencias

El modelo de la implementación es el Stochastic Gradient Descent (SGD) de la librería sklearn.

Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación.

Para la separación de los datos en un conjunto de entrenamiento y en un conjunto de validación se usó la función train_test_split de la librería sklearn.model_selection. Esta función recibe los datos de ***x*** y de ***y*** del dataframe a utilizar; como parámetros extras se le manda el tamaño del conjunto de pruebas, el cual en este caso fue del 20% total del dataframe. La función regresa los valores de ***x*** y de ***y*** para entrenamiento y validación.

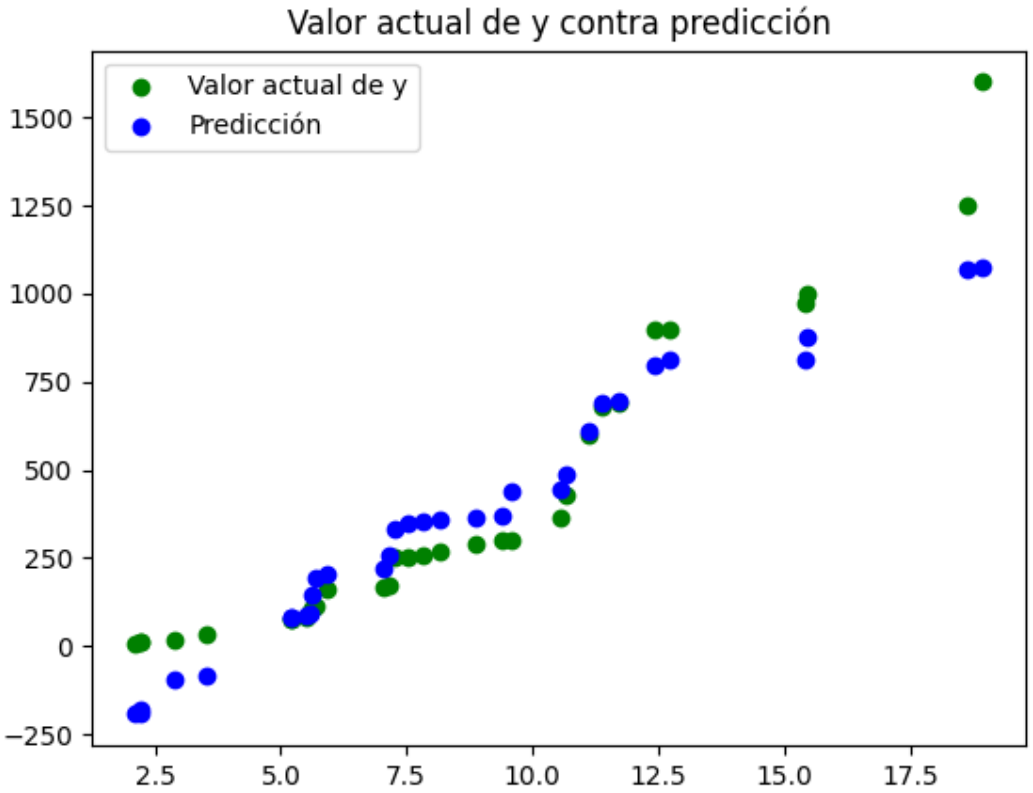
Una vez separados los datos se usa la función de StandardScaler de la librería sklearn.preprocessing para obtener mejores resultados en el modelo, esta función estandariza los valores del conjunto de datos. Se usa SGDRegressor para este modelo, el cual es de la librería sklearn.linear_model el cual recibe el número de épocas y el valor de alfa a utilizar. Se entrena el modelo con los valores de X_train y y_train.

Se usa el modelo para predicciones con los valores de X_test, se comparan las predicciones contra los valores de y_test.

Finalmente se obtiene el coeficiente de determinación de la predicción. Este coeficiente determina la calidad del modelo de regresión. Un valor de R^2 cerca de 1 indica que es un buen modelo de regresión. Para esto se usa la función score, la cual recibe X_test y y_test. El valor de R^2 para este modelo es de 0.8813.

Tabla comparativa de los valores de y_test y las predicciones del modelo.

Valor actual	Predicción
1250.0	1070.132185
80.0	90.337166
250.0	364.900703
975.0	878.548964
300.0	369.095924
690.0	697.757190



Diagnóstico y explicación el grado de bias o sesgo

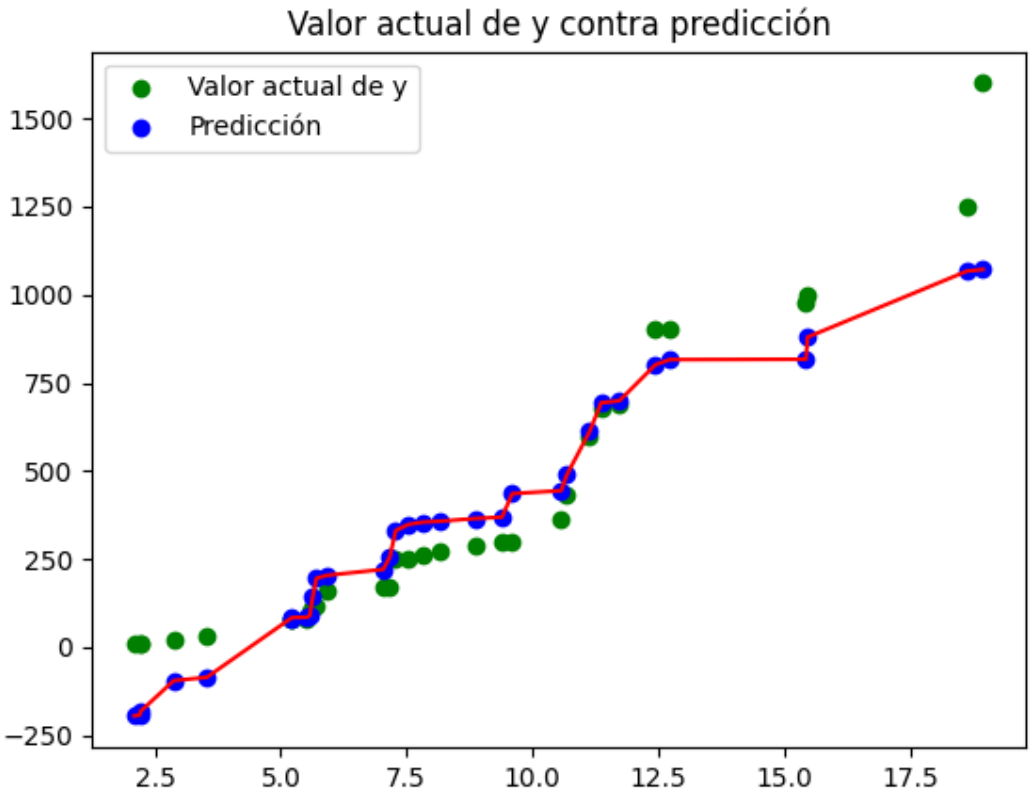
Para obtener el valor de bias lo que se hizo fue sacar el error cuadrático medio por medio de la función mean_squared_error que recibe y_test y y_pred dando un valor de 0.17. Lo siguiente es obtener el valor de la varianza, esto se hizo con la función de numpy np.var la cual recibe los valores de y_pred dando un valor de 0.94. Finalmente para obtener el bias se resta el error cuadrático medio menos la varianza, obteniendo en este caso un valor de -0.77. Al observar el valor de bias con relación al error cuadrático medio éste es un valor bajo, existe bastante distancia entre los valores reales y los predecidos. Se puede decir que el modelo entendió el patrón pero no la magnitud de los valores, por lo que el modelo no es preciso con valores reales. Esto se puede observar en la gráfica anterior.

Diagnóstico y explicación el grado de varianza

Para obtener el valor de varianza se utilizó la función de numpy np.var la cual recibe los valores de y_pred dando un valor de 0.94. Observando el valor de la varianza con relación al error cuadrático medio es un valor alto; el modelo funciona bien con valores de

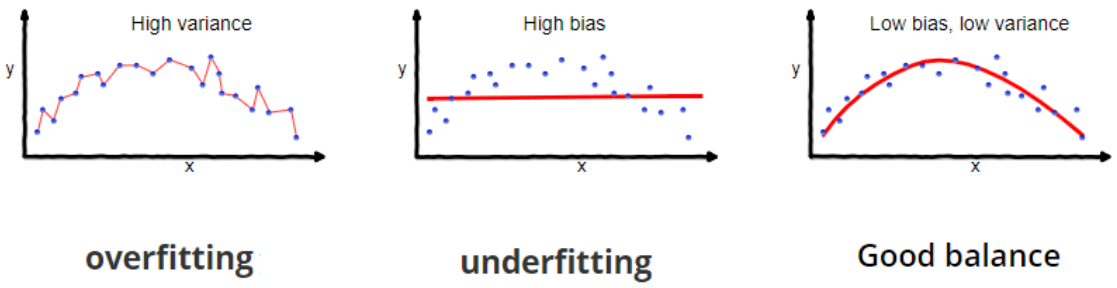
entrenamiento pero no es preciso con valores reales. El modelo tiene un enfoque en el ruido. Se puede asegurar que los valores utilizados para este modelo tienden a estar lejos de la media y tienen una alta variabilidad.

En la siguiente gráfica se observa el bajo bias y alta varianza por las curvas de la línea roja.



Diagnóstico y explicación el nivel de ajuste del modelo

El nivel de ajuste del modelo es overfitting, el primer indicador es el bajo bias y alta varianza y el otro indicador es la gráfica anterior que muestra la poca precisión con valores reales. Obteniendo muchas curvas en la línea que une cada predicción. A continuación se nuestra la comparación entre underfit, fit y overfit.



Mejoras para el desempeño del modelo

Una de las mejoras que se encontró es estandarizar los valores, haciendo esto se escalan los datos para que se ajusten a una distribución normal. Otra mejora encontrada es ajustar los hiperparámetros hasta obtener el mejor resultado posible. En este caso ajustando el valor de alfa a 0.1 y una cantidad de 1000 épocas el modelo obtuvo mejores resultados.

El error cuadrático medio obtuvo un valor de 0.15, la varianza 1.05 y el bias -0.90.