Tecnológico de Monterrey

Campus Querétaro

Asignatura

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

Tema

Módulo 2 Análisis y Reporte sobre el desempeño del modelo

Alumno

Carolina Arratia Camacho
A01367552

Fecha

12 de septiembre del 2023

Link al código que contiene los análisis hechos:

https://colab.research.google.com/drive/1SAQSX7u_GSE9NY_j8dnd291E07viD8VR?usp=sharing

El modelo que se utilizó para el análisis que se presenta en este apartado es el modelo realizado sin el uso de un framework.

Este busca que, a través de la implementación de un modelo de regresión lineal, y con ayuda de la función de gradiente descendiente, se pueda predecir el índice de felicidad de un país dado, tomando como indicadores las variables de Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity.

El dataset con el que se está trabajando cuenta con los datos del índice de felicidad de 158 países en diferentes años, contando con 315 filas. De estos el 90% están siendo usados para entrenar el modelo, y el otro 10% se usa para prueba.

Nota: Anteriormente se trabajaba solo con 158 datos, pero se decidió aumentar estos.

Posteriormente se realizó cross-validation dividiendo el train en 5.

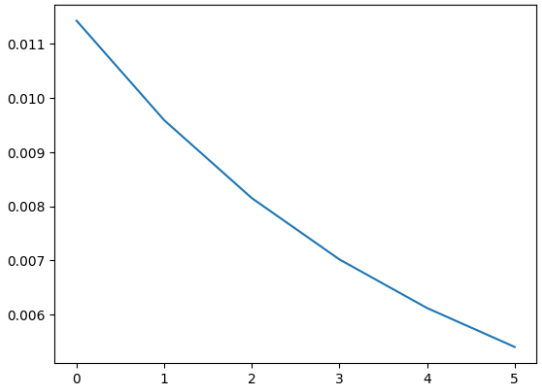
Dado que se trata de un modelo de regresión lineal, las métricas que se tomaron en cuenta para evaluarlo fueron la R^2 y el MSE.

La R^2 es útil ya que permite evaluar el rendimiento de mi modelo. Me dice qué tan bien mi modelo se ajusta a los datos, y la calidad de este para replicar los resultados. Po lo que, el obtener un valor de 1 sería lo deseado.

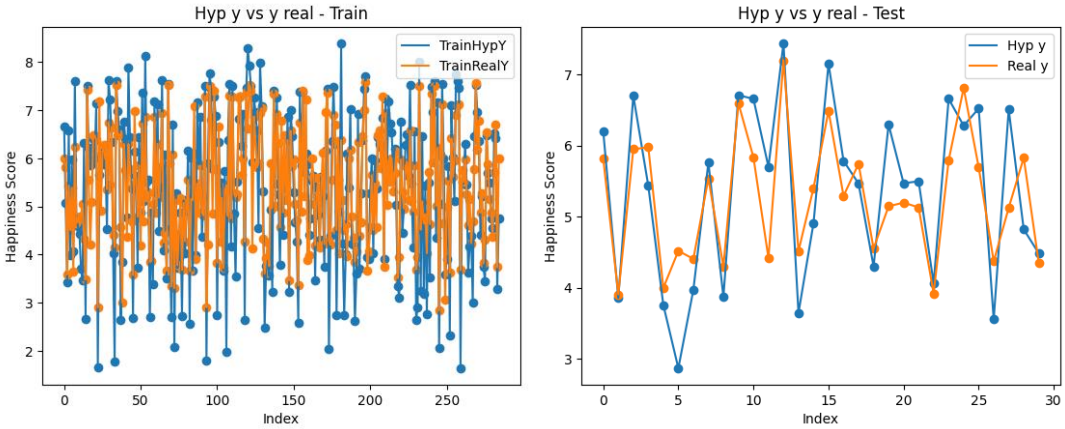
El Error Cuadrático Medio (MSE) también ayuda a medir la calidad del modelo ya que mide el promedio de los errores al cuadrado, es decir, la diferencia entre mi hipótesis de <y>, y su verdadero valor.

Datos obtenidos

Al correr mi modelo se obtuvieron diversas graficas.



Gráfica de loss para el entrenamiento



Gráficas comparativas de los valores reales y predichos tanto en el test como en el train

Asimismo, se realizó el cálculo del coeficiente de determinación o R^2, con el fin evaluar el desempeño de modelo, y de cierta manera determinar su accuracy. Con este cálculo se obtuvieron los siguientes datos:

R^2 de Train 0.4323

```
R2 score for train 0.43231231709558904
```

R^2 de Test 0.3363

```
R2 score for test 0.3363550103954128
```

Promedio de R^2 obtenidas de Cross-Validation aplicado al Train 0.34

Además, al calcular el Error Cuadrático Medio siguiendo la fórmula de

$$MSE = (real - estimado)^2$$

se obtuvo:

MSE de Train: 0.7680

MSE de Test: 0.5113

Análisis

Una vez teniendo los datos anteriores se puede observar que existe un gran sesgo en cuanto a los datos predichos y los reales. Algo que se ve más claro en las gráfica de “Hyp y vs y real - Train” y de “Hyp y vs y real - Test”, dado que hay muchos puntos azules que no coinciden con los naranjas. Esto se puede deber a la variabilidad de los datos, por lo que el modelo presenta un alto bias o sesgo. Esto también se puede observar al ver los valores de R^2, al tener valores de R^2 tan pequeños se infiere que el modelo no está teniendo un buen ajuste y por ende, los valores predichos difieren de manera significativa de los reales.

También se nota un Error Cuadrático medio alto que de la mano con lo ya antes mencionado sobre el sesgo que existe entre los datos predichos y los datos reales.

Otro punto a mencionar es que existe una diferencia del 0.10 aproximadamente entre mi coeficiente de R^2 del Train y del Test, siendo mi coeficiente en el Train mejor, pero no se presenta una diferencia muy significativa entre estos.

A la hora de aplicar el Cross-validation se notó que, en cada una de las iteraciones, se realizaron 5, se obtuvo un valor muy similar a 0.33, por lo que se considera que presenta una varianza baja, al ser similar a mis demás valores de R^2.

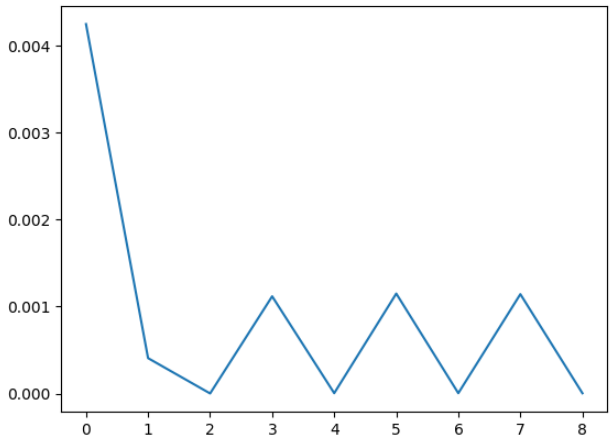
En cuanto al modelo, este presenta underfitting, el modelo no es capaz de encontrar la relación entre las variables dependientes y mi variable dependiente, que es la que se busca predecir. Dado que es un modelo simple, el bias presentado es elevado.

Mejoramiento del modelo

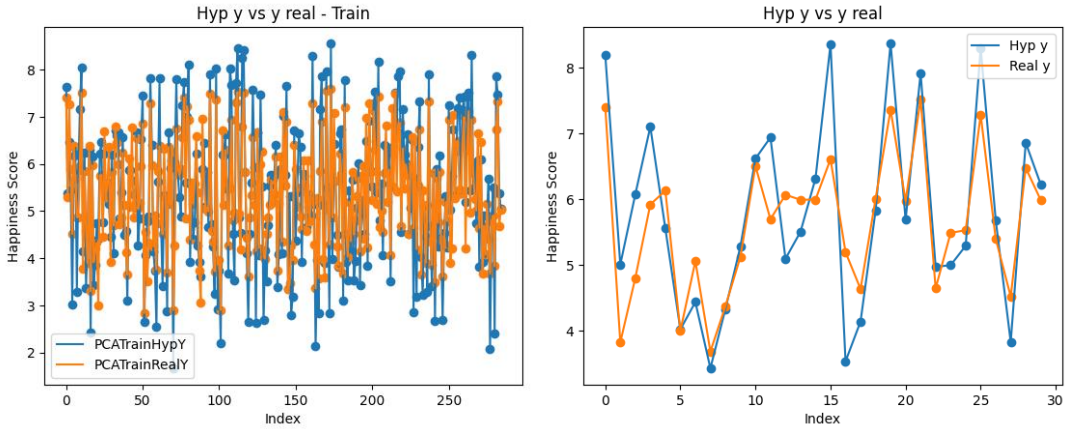
Con las pruebas anteriores se determinó que el modelo no era capaz de representar la relación entre las variables independientes y la dependente, y por ende no es muy preciso. Teniendo un rendimiento poco eficiente tanto en los datos de training como en los de test.

Con el propósito de mejorar el modelo se decidió hacer uso del Principal Component Analysis. Dado que es muy complicado el encontrar la relación entre las 6 variables independientes y la variable dependiente, al hacer uso del PCA permite encontrar las variables más significativas y así trabajar con menos variables.

Las gráficas obtenidas con esta mejora son las siguientes



Gráfica de error del entrenamiento



Gráficas comparativas de los valores reales y predichos tanto en el test como en el train

Los valores de R^2 obtenidos de este nuevo modelo son:

R^2 de Train 0.5495

```
r2 score for perfect model is 0.5495046594250936
```

R^2 de Test 0.4327

```
r2 score for perfect model is 0.4327277406704626
```

Mientras que sus valores de MSE son

MSE Train: 0.5922

MSE Test: 0.6092

Al analizar la R^2 de este nuevo modelo se obtiene que el tanto en el train como el test, este valor ha subido, es decir, ha mejorado. La R^2 de Train pasó de 0.4323 a 0.5495. Mientras que la de Test pasó de 0.3363 a 0.4327.

Recordando que lo ideal es que estos coeficientes tuvieran un valor de 1 ya que indicarían que mi modelo se está ajustando a los datos. A pesar de que aún están bastante lejos de alcanzar ese ideal, sí hubo una mejora. Además, recordando que al ser aplicado el PCA se necesitan menos recursos para correr el modelo ya que se vuelve menos complejo que cuando tenía todas mis variables independientes en mi modelo inicial.