

Análisis y reporte sobre el desempeño del modelo

Keyuan Zhao A01366831

Resumen

El objetivo del presente documento es explicar el contexto del dataset y el modelo de predicción utilizado, además, mostrar el análisis del modelo establecido para determinar si es útil para el dataset.

Introducción

La ubicación geográfica de Australia, el clima está muy influido por las corrientes oceánicas que correlaciona con las sequías periódicas y las estaciones se invierten. Por esta razón, los meses más fríos son en verano, mientras que los más cálidos son en invierno.

Contexto del dataset y regresión lineal

El dataset que se usó para implementar el modelo fue registrado por el gobierno de Australia por más de diez años, en donde contiene datos principales como temperatura mínima y máxima de cada ciudad dividido por días, y datos secundarios como velocidad y la dirección del viento, humedad, lluvia, evaporación y entre otros datos que construye dicha dataset.

Antes de iniciar el modelado se necesita verificar si los datos son adecuados para el uso, es decir, el dataset seleccionado no debe de haber datos nulos (Nan). Por lo tanto, el primer paso es usar la función de dropna para eliminar la fila donde contiene los datos nulos.

Para facilitar la implementación del modelo, sólo se usó los datos de temperatura mínima y máxima de la ciudad Canberra (capital de Australia) para predecir la temperatura con la función de regresión lineal simple, en donde "x" es temperatura mínima y "y" es la temperatura máxima. Al conocer estas variables, podemos usar la regresión lineal simple con la ayuda de la librería sklearn, ya que es fácil para predecir el modelo según con los datos dependientes e independientes (Figura 1).

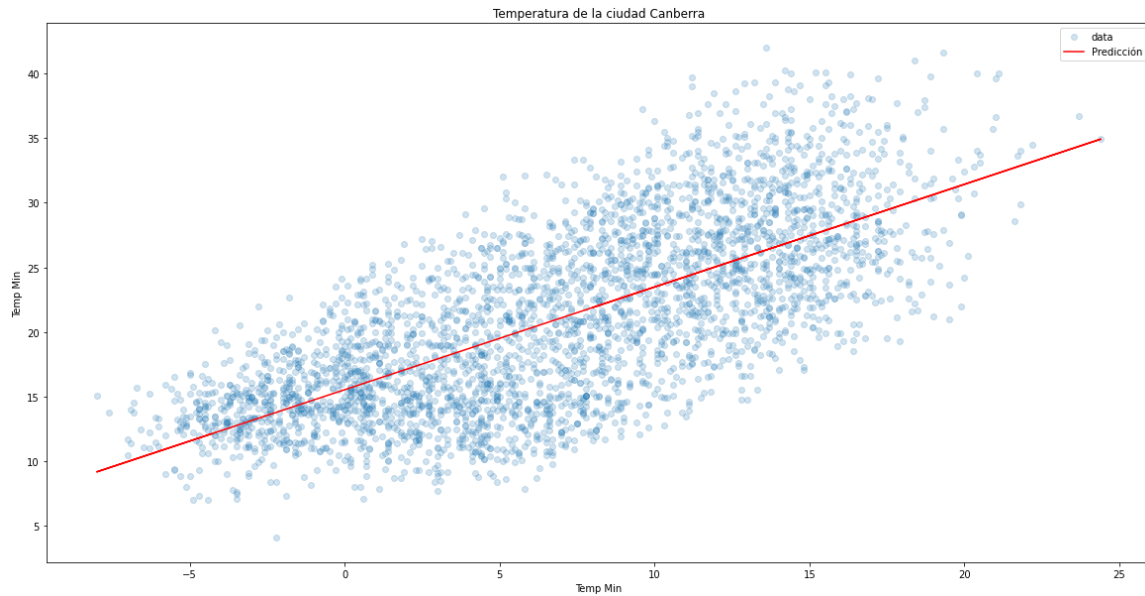


Figura 1. Gráfica de temperatura máxima vs. temperatura mínima de Canberra

Así como se muestra en la figura 1, los puntos azules son los datos del dataset y la línea roja indica la predicción del modelo después de usar la librería sklearn. A simple vista los datos son muy dispersos rodeando la línea de regresión, entre más cerca de la línea hay más datos concentrados.

Análisis del modelo

Para tener un análisis profundo, el primer paso fue dividir los datos del dataset en dos partes: train con la 2/3 parte y test con 1/3 parte. Ya que esto es fundamental para poder analizar con la técnica de cross validations.

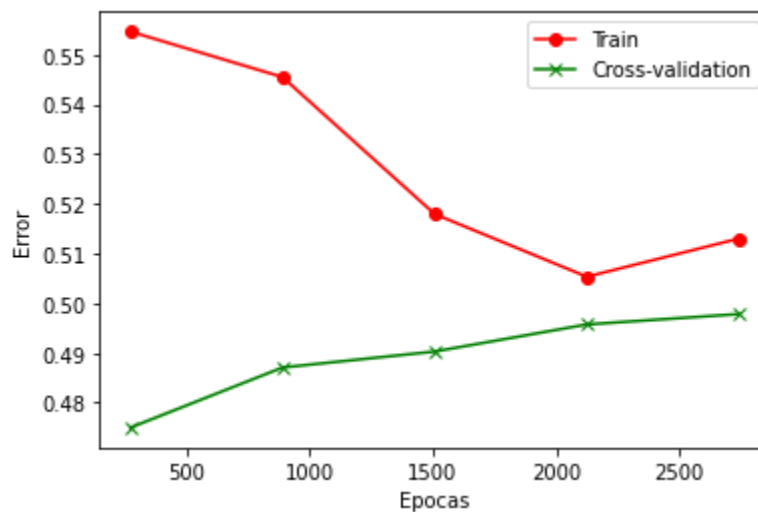


Figura 2. Gráfica de error vs. épocas (train y cross validation)

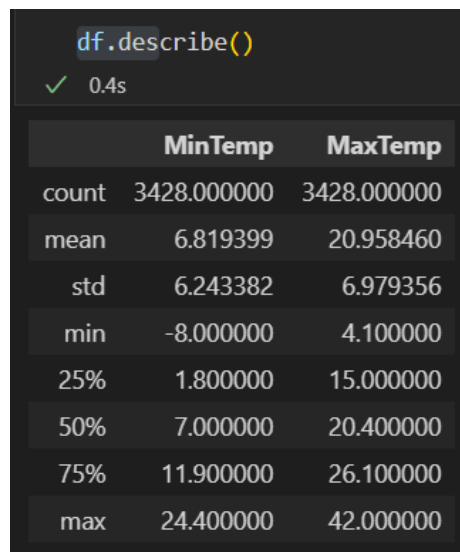
En la figura 2 podemos ver que el modelo tiene un grado de bias o sesgos muy alto, ya que los modelos de regresión lineal son rápidos y fácil de aprender para las máquinas, pero a su vez son menos flexibles debido a que la predicción es lineal. Esto hace que la línea no abarque la gran cantidad de datos por su dispersión, es decir, el rendimiento predictivo es menor debido a la complejidad de datos.

El modelo tiene el grado de varianza baja debido a que es un modelo paramétrico que presenta con poca flexibilidad; además, los datos que representa la temperatura no varían mucho si hace modificaciones ya que se encuentra estable.

Al recopilar el grado de bias o sesgos y el grado de varianza, podemos determinar que el modelo es underfitting, ya que el método utilizado para establecer el modelo es muy simple para esta gran cantidad de datos, por lo cual se necesita utilizar otras formas u otros algoritmos de Machine Learning para mantener el balance bias y la varianza. alguna de estas formas puede ser cambiar el modelo a regresión polinomial y usar datos más significativos, es decir, usar la temperatura promedio como otra variable para poder predecir la temperatura del siguiente día.

Refinamiento del modelo

Para balancear el grado de bias y varianza, debemos analizar de nuevo los datos del dataset para seleccionar los valores que más se representan. Por lo tanto se realizó la descripción del dataset:



```
df.describe()
```

	MinTemp	MaxTemp
count	3428.000000	3428.000000
mean	6.819399	20.958460
std	6.243382	6.979356
min	-8.000000	4.100000
25%	1.800000	15.000000
50%	7.000000	20.400000
75%	11.900000	26.100000
max	24.400000	42.000000

Figura 3. Descripción del dataset

En la figura 3 podemos ver que existe una gran diferencia entre el mínimo de los datos y el 25% de los datos, esto también pasa con el máximo de los datos con el 75% de los datos,

es decir, que hay muchos datos con valores extremos afectando al modelo. Por lo tanto, el proceso de refinamiento fue eliminar esos datos seleccionando sólo 25% - 75% de los datos más representativos del dataset.

```
df['MinTemp'] = df['MinTemp'][(df['MinTemp'] >= 1.8) & (df['MinTemp'] <= 11.9)]  
df['MaxTemp'] = df['MaxTemp'][(df['MaxTemp'] >= 15) & (df['MaxTemp'] <= 26.1)]  
df = df.dropna()
```

Figura 3. Selección de datos

Después, se aplicó el mismo procedimiento: dividiendo en train y test y aplicar la técnica de cross validation, y arrojo el siguiente resultado:

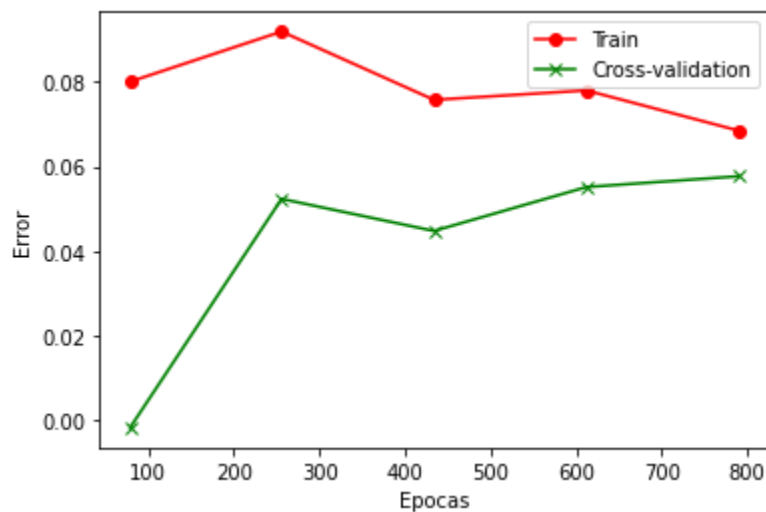


Figura 4. Gráfica del modelo refinado

En la figura 4 podemos notar que el grado de bias se redujo cierta manera después de descartar los valores extremos, por lo cual el nuevo grado de bias es medio y el grado de varianza se mantuvo en bajo. Con estos resultados podemos considerar que, el modelo todavía se necesita bajar un grado de bias para ser una modelo óptima, por lo cual, todavía se necesita hacer refinamiento sobre el modelo.