04-11-2022

Resumen Ejecutivo.

Telemetry Sensor Data.

Jorge Esneider Henao Gonzalez.

La telemetría para IOT es en este momento una de las tecnologías disruptivas que ha traído la industria 4.0 con más auge, razón por la cual, en este trabajo se desarrolla un proceso de recolección de datos de lecturas de variables tales como: temperatura, humedad, monóxido de carbono (CO), gas licuado de petróleo (GLP), humo, luz y movimiento, con el fin de recolectar datos en tres distintas condiciones. Este proceso de recolección de datos se realiza, a través sensores que permiten leer con el mismo protocolo de comunicación en una raspberry pi, quien, a su vez, permite recolectar la data, en las tres condiciones anteriormente mencionadas, en pro de realizar un modelo de Machine Learning.

device	environmental conditions
00:0f:00:70:91:0a	condiciones estables, más frescas y más húmedas
1c:bf:ce:15:ec:4d	temperatura y humedad muy variables
b8:27:eb:bf:9d:51	condiciones estables, más cálidas y secas

Tabla 1. Condiciones de datos.

Los datos fueron obtenidos a partir de una serie de conjuntos de sensores idénticos, hechos a medida, mismos que fueron conectados a una Raspberry Pi y estos dispositivos IOT se instalaron para tomar datos en 3 diferentes ubicaciones físicas con diversas condiciones ambientales.

Estas variables llegan a la raspberry pi en los siguientes tipos de datos.

1. ts (timestamp) ==> epoch
2. device id ==> object
3. CO (Carbon Monoxide) in ppm ==> float64
4. humidity in percent ==> float64
5. light ==> bool

6. LPG (liquified Petroleum Gas) in ppm ==> float64
7. motion ==> bool
8. smoke in ppm ==> float64
9. Temperature in Fahrenheit ==> float64

Proceso.

Una vez recolectados los datos, se procede a realizar un pre – procesamiento, logrando analizar los datos por medio de gráficas, hisogramas y boxplot[1], identificando que existen una amplia correlación de datos, en algunas variables tales como motion, lpg, smoke, co; luego, se logró identificar que el modelo de las variables temperatura, humidity, light, son las variables que van a permitir predecir en que condición se encuentra el prototipo, es decir: condiciones estables, más frescas y más húmedas; temperatura y humedad muy variable o condiciones estables, mas cálidas y secas; resultado que se verá reflejado en los números 0, 1 y 2, respectivamente.

El primer modelo que se implementa al momento del procesamiento de los modelo de regresión lineal, se realizó un experimento, que consistió en trabajar con estandarización y sin uso de la estandarización en pro de revisar su proceso y su comportamiento y aumentar experiencia en los procesos; luego se tomó la decisión de trabajar con el modelo sin realizar estandarización conozco Coeficiente de Determinación R2 = 0.8232507139737062 y finalmente, se desarrolló un código que permitió realizar una predicción, obteniendo un acierto de más o menos el 82%.

Luego se procedió a realizar el proceso con el modelo de Lasso y obtuvo el Coeficiente de Determinación R2 = 0.81, en el que se descartó este modelo frente a la mejor respuesta que se obtuvo con el modelo de regresión lineal.

El modelo de Ridge se implementó y se obtuvo una medida de coeficiente de determinación muy idéntico a la regresión lineal, razón por la cual se realizó el proceso de encontrar el alfa, obteniendo un 6.1 y posteriormente se realizó cross validation para ridge obteniendo de nuevo valores muy cercanos a el modelo de regresión lineal.

Conclusiones.

Teniendo en cuenta lo anteriormente expuesto, se logra obtener un modelo de predicción tanto para ridge y linear regresión, en el que ambos obtuvieron un comportamiento muy similar.

El coeficiente de determinación tanto para ridge como para linear regresión son muy idénticos

[1] Boxplot: Diagrama de caja.

Tanto el Error Cuadrático Medio (Mean Squared Error - MSE) como la Raíz Cuadrada del Error Cuadrático Medio de ridge como de linear regresión fueron de igualmente cercano.

Los coeficientes estimados por ridge regression pueden verse alterados si las variables no se estandarizan antes de llevar a cabo el ajuste.

La ventaja con la que cuenta el ridge regression respecto al método de mínimos cuadrados reside en el equilibrio bias-varianza: conforme λ aumenta, la flexibilidad del ajuste por ridge regression disminuye, lo cual disminuye la varianza, pero aumenta el bias.

Un pequeño cambio en el set de datos de entrenamiento puede hacer variar los coeficientes de manera sustancial en línear regresión, mientras que ridge regression puede reducir considerablemente la varianza a expensas de un pequeño aumento en el bias (conforme λ aumenta).