Momento de Retroalimentación: Módulo 2 Implementación de una técnica de aprendizaje máquina sin el uso de un framework. (Portafolio Implementación).

Fabián Erubiel Rojas Yáñez, A01706636
Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Querétaro, México
A01706636@tec.mx

Resumen - En este proyecto se aborda el desarrollo de una técnica de aprendizaje máquina sin el uso de frameworks. En esta implementación abordaremos el uso de una regresión logística para la clasificación de una variable binaria, en este caso y debido a la naturaleza de mi dataset, se usa como predicción para una estrategia de marketing de un banco, en el que se busca predecir si un cliente abrirá un deposito con el banco despues de algunas llamadas por parte de la institución.

I. INTRODUCCIÓN.

La regresión logística es una técnica de machine learning o aprendizaje máquina, utilizada para la predicción o clasificación de problemas lineales. Para lograr hacer esto usa funciones como la "sigmoide", que se encarga de mapear valores y asignarlos a un rango probabilístico, lo que permite clasificar de una excelente forma valores entre 0 y 1, por eso en clasificación binaria es una excelente técnica.

En este trabajo, se presenta la implementación de una regresión logística desde 0, sin usar frameworks o librerías dedicadas a machine learning. Esto con el enfoque de desarrollar un mejor entendimiento de los modelos y posibles aplicaciones de machine learning. También aporta una comprensión sólida de técnicas de clasificación.

El dataset y la problemática a tratar es acerca de una campaña de marketing de un banco portugues, que busca predecir si un cliente se suscribirá o no a un depósito de plazo fijo, este acercamiento, normalmente es por llamadas telefónicas y dentro del dataset se incluyen datos acerca de estas mismas, como la duración y los contactos realizados posteriormente (si son mayores a 1).

II. Dataset y su división..

Dentro de el dataset escogido, se muestran los siguientes variables dentro del dataset:

Las variables categóricas y numéricas detectadas son las siguientes:

Las variables numéricas son aquellas que tienen un tipo de dato int64 y representan cantidades o medidas que se pueden ordenar y operar matemáticamente.

- age: Edad del cliente (int64)
- balance: Saldo promedio anual en euros (int64)
- day: Día del mes en que se realizó el último contacto (int64)
- duration: Duración del último contacto en segundos (int64)
- campaign: Número de contactos realizados durante esta campaña (int64)
- pdays: Número de días desde el último contacto en una campaña anterior (int64)
- previous: Número de contactos realizados antes de esta campaña (int64)
- Variables Categóricas

Las variables categóricas son aquellas que tienen un tipo de dato object y - representan categorías o etiquetas que no tienen un orden inherente.

- job: Tipo de trabajo (object)
- marital: Estado civil (object)
- education: Nivel educativo (object)
- default: Si tiene crédito en incumplimiento (sí/no) (object)
- housing: Si tiene préstamo hipotecario (sí/no) (object)
- loan: Si tiene préstamo personal (sí/no) (object)
- contact: Tipo de contacto (object)
- month: Mes del último contacto (object)
- poutcome: Resultado de la campaña anterior (object)
- y: Si el cliente se suscribió a un depósito a plazo fijo (sí/no) (object) - (variable objetivo)

Preprocesamiento del dataset:

Para hacer el procesamiento o "limpieza" del dataset usado, se usaron varias técnicas como el eliminar valores nulos o vacíos, valores duplicados valores "outliers" que son valores que al ser muy variables dentro del rango de valores obtenidos pueden afectar de forma negativa a el modelo, para hacer esto se usaron gráficas de caja, para establecer unos valores "medio" y detectar variables fuera de estos rangos medios.

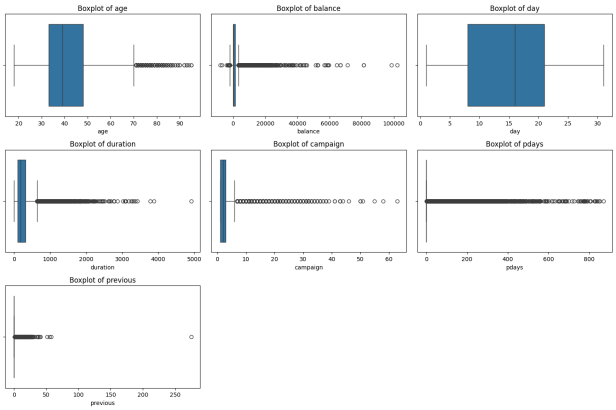

Fig 1. Gráficos de caja para las variables numéricas.

Para poder eliminar estos outliers, use la técnica del valor intercuartílico (IQR), que básicamente es una medida que ayuda a entender la dispersión de los datos entre dos distintos cuartiles Q1 Y Q3:

- Q1 = Es el primer cuartil, de el valor por debajo del 25%.
- Q3 = Es el tercer cuartil, de el valor que esté por debajo del 75%.

Siguiendo la siguiente fórmula:

$$IQR = Q3 - Q1$$

El resultado de esto nos da una rango central donde se encuentren los datos de alrededor del 50%, los demás datos que están fuera de este rango podrían afectar negativamente modelos que posteriormente se implementaran, ya que son valores o muy grandes o muy pequeños que podrían causar un ruido que se puede evitar.

- Límite Inferior: Se define como Q1 - 1.5 * IQR. Cualquier valor por debajo de este límite se considera un outlier inferior.
- Límite Superior: Se define como Q3 + 1.5 * IQR. Cualquier valor por encima de este límite se considera un outlier superior.

Posteriormente procedí a eliminar columnas que considero son irrelevantes para el modelo:

- day
- month
- pdays
- previous
- poutcome
- contact

Considero que no tienen demasiada relevancia para la predicción que busco hacer.

Despues procedi a disminuir las variables de trabajo de 12 a solamente 7,

```
job_map = {
    'management': 'Management',
    'admin.': 'Management',
    'entrepreneur': 'Management',
    'technician': 'Technical',
    'blue-collar': 'Technical',
    'services': 'Technical',
    'self-employed': 'Technical',
    'housemaid': 'Unskilled',
    'unemployed': 'Unemployed',
    'retired': 'Unemployed',
    'student': 'Student',
    'unknown': 'Other'
}
```

Esto con la finalidad de agrupar datos de la columna de trabajo y al momento de trabajar con ellas, sea más sencillo al tener menos variables disponibles. Esto debido a que se puede simplificar por el hecho de que no es tan relevante y podría hacerse más, puesto que sería más sencillo si se sabe que se trabaja o no, ya que esto afecta en la economía o aceptación del plazo con el banco.

Después, aplique la técnica de One Hot Encoding para las variables categóricas y convertirlas a variables binarias, esto con el fin de trabajarlas de forma numérica (binaria), ya que éstas considero que sí son relevantes en el analisi de la información y el desarrollo del modelo, por lo que trabajarlas de forma numérica me resulta más sencillo.

También use la técnica de escalamiento o normalización de datos, para los datos numéricos, ya que quería manejarlos de forma numérica y que fuera más sencillo para mi. Y que las grandes diferencias entre valores de las variables no afectarán negativamente al modelo.

III. Regresión Logística.

Elegí la regresión logística para abordar este problema, ya que se busca predecir si un cliente se suscribe o no a la institución bancaria, es decir es un problema binario que se reduce a “si” o ”no”. Para este tipo de problemas la regresión logística es muy eficiente y precisa, también es más sencilla de implementar que una red neuronal, al ser un problema lineal, si se puede ver de cierta forma una red neuronal es un conjunto de regresiones logísticas.

Use la función sigmoide de la regresión logística para mapear los valores entre 0 y 1, que básicamente es fundamental y el objetivo de la problemática, esto para arrojar valores lo más cercano posibles a 0 o 1.

También use la función de Cross Entropy para hacer los cálculos de la diferencia de los valores reales y los predecidos, esta función guía al modelo de aprendizaje para aprender de forma “Correcta” predecir o clasificar los datos.

De igual forma hice uso del algoritmo de Gradiente

descendiente para disminuir el error del modelo, en el entrenamiento del modelo de regresión, más en concreto desglosar este algoritmo, ya que es uno de los más importantes de mi modelo:

En concreto lo que busco de este algoritmo es ajustar los pesos y el bias para disminuir el error o el costo del modelo, esto para que las predicciones se acerquen más a las etiquetas deseadas y hacer preciso el modelo. Sirve para monitorear cómo va mejorando el algoritmo tras cada iteración o grupo de iteraciones.

Use una funcion de prediccion para establecer las salidas de cada valor, es decir predice si es “1” o “0”. Esta se alimenta con las entradas, el bias y los pesos, en base a un calculo realizado con ellos y usando la fucnion sigmoide es que se establecen los valores si son “1” o “0”.

IV. Resultados obtenidos.

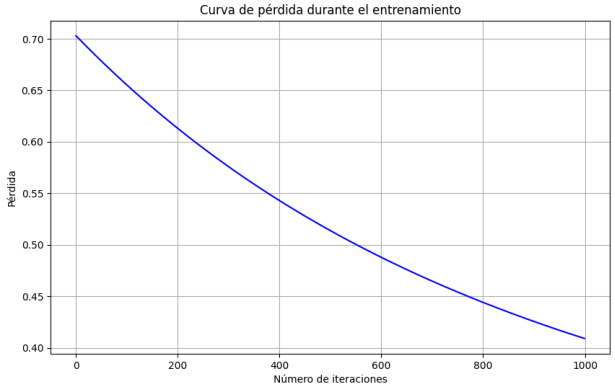

Fig 2. En esta gráfica se puede observar la pérdida durante el entrenamiento.

La interpretación de esta gráfica, quiere decir que el modelo va aprendiendo cada vez que hay más iteraciones, por lo que podemos deducir que va ajustando los parámetros de forma correcta para lograr aumentar su precisión de las predicciones. Por la forma de la curva de esta gráfica se puede deducir que no se observa overfitting y al ser tan suave, también puedo deducir que el modelo es estable y bueno.