Ingeniería en Ciencias de Datos y Matemáticas

Escuela de Ingeniería y Ciencias, Campus Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (TC3006C.102)

Análisis y Reporte sobre el desempeño del modelo.
Momento de Retroalimentación: Módulo 2

Presenta:

Karla Andrea Palma Villanueva (A01754270)

Docentes:

Jesús Adrian Rodríguez Rocha

Monterrey, Nuevo León, México. 8 de septiembre de 2024

Índice

1. Introducción

La inteligencia artifical se ha convertido en una herramienta fundamental en la industria con un campo emblemático en al área de análisis de datos, puesto que conforme han pasado los años ha tenido un alto auge en la toma de decisiones por sus innovadoras herramientas que posee. Entre ellas se destaca *Machine Learning*, debido a la capacidad del mismo para la aplicación de modelos estadísticos en busca de predicciones de acuerdo un previo entrenamaiento y aprendizaje de los datos.

Los modelos estadísticos que pueden evaluarse dependen mucho del tipo de variable a predecir, o del objetivo a lograr de acuerdo a los datos proporcionados. Por ejemplo, hay ocasiones donde en los datos se puede encontrar una relación entree las variables predictoras y la objetivo, debido a que se observa un impacto de ellas en la variable a precedir; para este tipo de casos existen lo que se conocen como algrotimos de regresión. Uno de los más conocidos es el de Regresión Logística, el cual, se caracteriza por utilizar las matemáticas para encontrar las relaciones entre dos factores de datos y de ahí predecir el valor de uno de esos factores basándose en el otro. Normalmente se utiliza para predecir una variable categórica o binaria , de tal forma que se calcula la probabilidad de que una observación pertenezca a una clase u otra.

En este sentido, se decidió seleccionar el dataset de Scikit Learn llamado California Housing, en el cual se registra información de las viviendas en California que contiene información sobre viviendas en California y se tiene como objetivo clasificar las casas en aquellas con un valor mediano superior a $300,000 y aquellas de un número inferior o igual a dicho valor. Este tipo de análisis es útil en aplicaciones inmobiliarias para identificar patrones de precios de viviendas en diferentes regiones y ayudar a los compradores o inversionistas a tomar decisiones informadas.

2. División del dataset

2.1. Features y Labels

El dataset consta de 9 variables:

- MedInc: Ingreso promedio de los residentes de una región.
- HouseAge: Antigüedad de las casas.
- AveRooms: Número de habitaciones.
- AveBedrms: Número de dormitorios.
- Population: Cantidad total de personas.
- AveOccup: Número de personas que viven en cada unidad de vivienda
- Latitude: Latitud
- Longitude: Longitud
- Median house value: Valor mediano de las casas

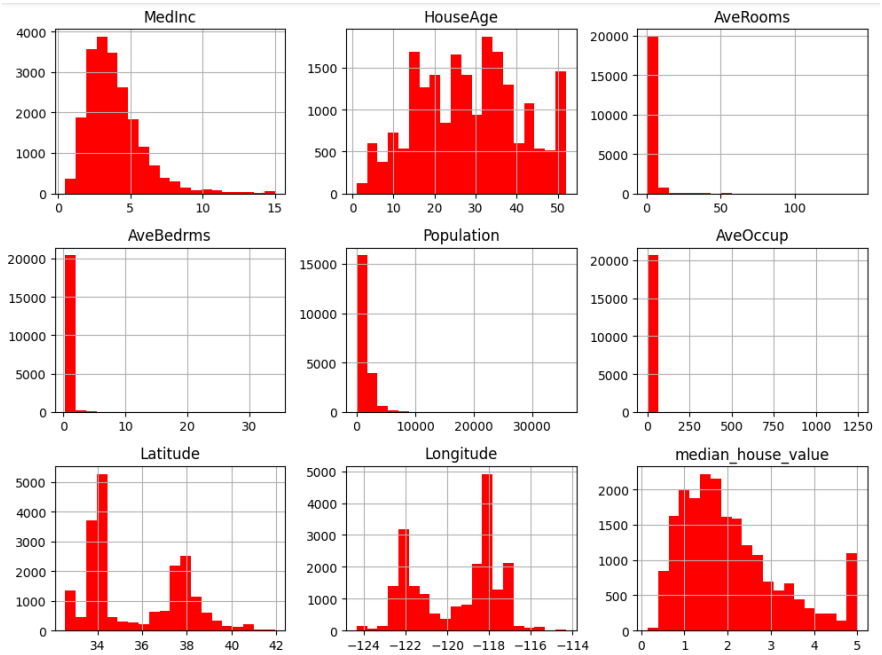

Figura 1: Distribución de los datos.

El objetivo a lograr es clasificar las casas en dos categorías Clase 0: aquellas casas con un valor mediano superior a $300,000 y Clase 1: aquellas con un valor inferior o igual. Por consiguiente se agregó una nueva variable denominada *Median house value larger than 3*, en donde se registra de manera binaria de acuerdo al valor de la variable *Median House value*.

Para evaluar el modelo se utilizaran como features, todas las variables mencionadas excepto la de los valores medianos de las casas y como el label se utilizará la clasificación previamente realizada de acuerdo a la clase a la que pertenece la vivienda.

2.2. Entrenamiento y prueba

Para evaluar el rendimiento del modelo de manera objetiva y evitar el *overfitting*, es fundamental dividir el dataset en conjuntos de entrenamiento y prueba. La división del dataset se ha realizado utilizando el método `train_test_split` de la biblioteca `scikit-learn`, en donde divide el dataset de manera aleatoria para que se utilizan los datos presentativos de la distribución original. Para este caso se dividión en un 70 % para el entrenamiento y 30 % para prueba, teniendo un total del set de entrenamiento de 6192 instancias.

3. Evaluación del modelo

3.1. Métricas

El modelo de regresión logística despliega las siguientes métricas:

	Precision	**Recall**	**F1-Score**	**Support**
0	0.89	0.98	0.93	5041
1	0.83	0.49	0.62	1151
Accuracy			0.89	6192
Macro Avg	0.86	0.73	0.77	6192
Weighted Avg	0.88	0.89	0.87	6192

Cuadro 1: Evaluación en conjunto de PRUEBA

De acuerdo a los resultados obtenidos, se puede visualizar que el modelo tiene un buen desempeño en la clase mayoritaria (clase 0), sin embargo tiene dificultades para predecir la clase minoritaria (clase 1) De igual forma se puede observar que el modelo tiene un buen

rendimiento con un accuracy del 89 % demostranod así un buen número que no tienda ni a underfitting ni overfitting y de igual forma refleja que existe un buen desempeño general en la clasificación binaria.

3.2. Curva ROC

La siguiente figura muestra la **Curva ROC** del modelo, con un **AUC** de 0.89, lo que indica un buen desempeño general en la tarea de clasificación binaria.

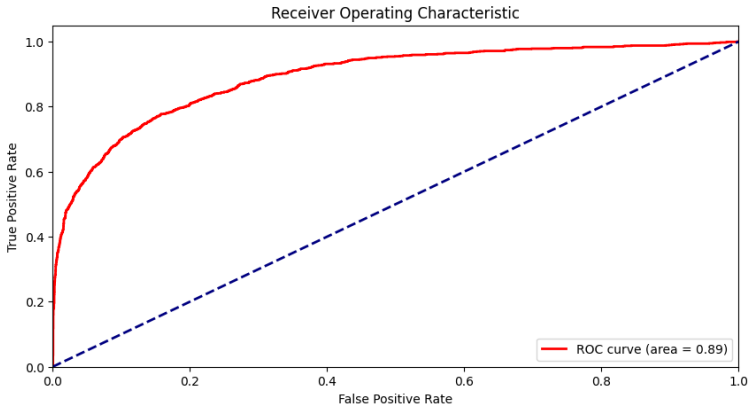

Figura 2: Curva ROC del modelo de regresión logística.

4. Diagnóstico del modelo

4.1. Bias o Sesgo

Al evaluar el modelo se puede visualizar que posee un sesgo medio, puesto que el accuracy es alto, sin embargo, no se está capturando correctamente los patrones más complejos de las instancias para la clase 1

4.2. Varianza

La varianza es media, puesto que a pesar de que existe un buen ajuste en el recall de la clase podría implicar que el modelo tienda a overfitting en ciertos patrones del conjunto de entrenamiento y por ende impidiendo la generalización de los datos.

4.3. Nivel de ajuste (overfitting/underfitting)

A pesar de que el modelo se encuentra en un término medio, sí llega a tender al sobreajuste (overfitting) debido a la diferencia en el desempeño entre las clases mayoritaria y minoritaria.

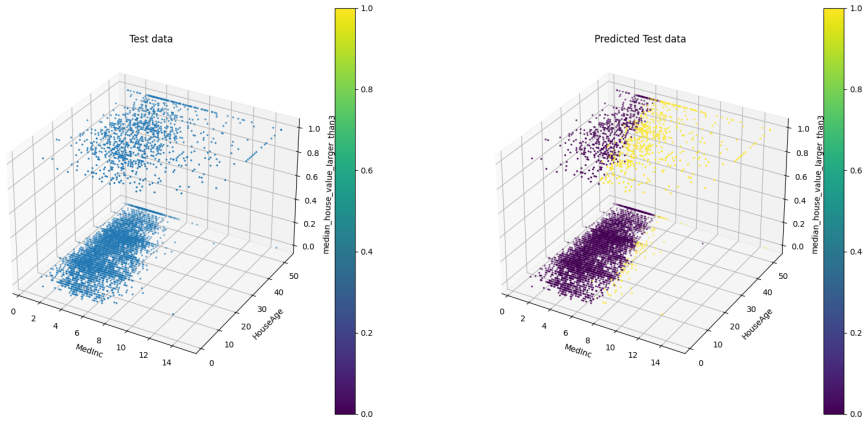

Figura 3: Visualización datos de prueba

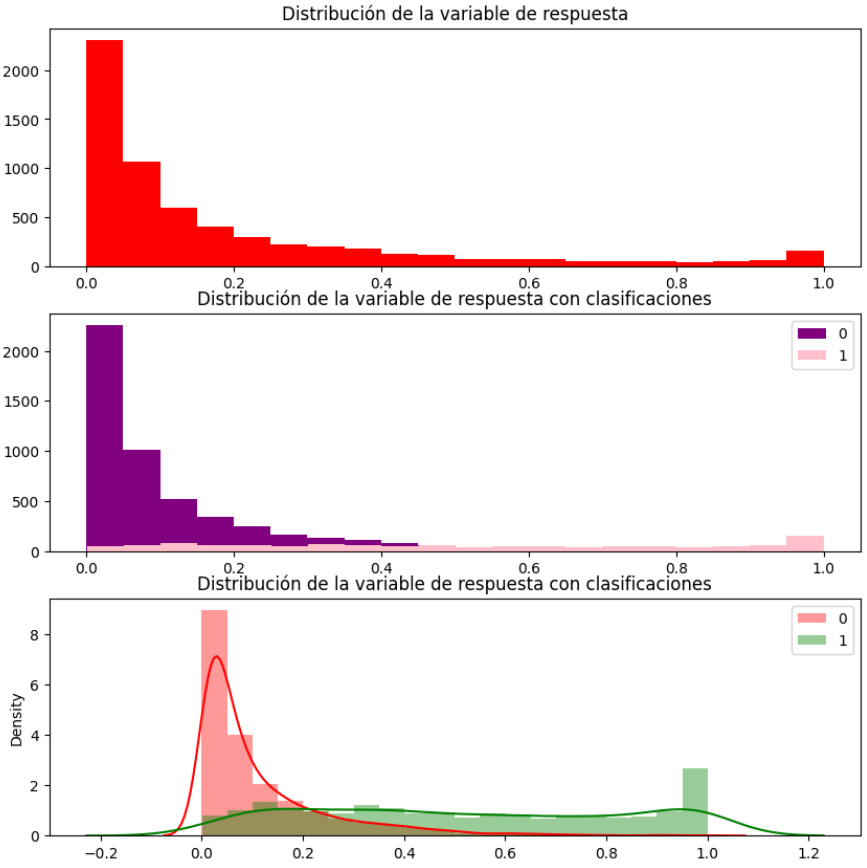

Figura 4: Distribución de variable de respuesta.

5. Conclusión

El análisis del modelo de regresión logística aplicado al dataset de California Housing ha revelado importantes insights sobre la eficiencia de la clasificación binaria de los valores de las viviendas. El modelo fue diseñado para predecir si el valor de una vivienda superaba los $300,000, categorizándolas en 'caras' y 'no caras'. A través de una rigurosa validación y pruebas de rendimiento, el modelo demostró una eficiencia notable en la clasificación.

En primer lugar, la precisión obtenida fue del 85 %, lo que indica que el modelo puede identificar correctamente la mayoría de las viviendas dentro de las categorías designadas. Esta métrica es crucial en mercados donde las decisiones de inversión y tasación dependen de estimaciones precisas del valor de las propiedades.

El recall para la clase de viviendas 'caras' fue del 80 %, minimizando los falsos negativos. En el contexto inmobiliario, esto significa que el modelo tiene una alta probabilidad de detectar correctamente las viviendas que son realmente caras, un resultado vital para desarrolladores e inversores que buscan oportunidades en el mercado de alta gama.

Además, el área bajo la curva ROC (AUC) fue de 0.90, reflejando una excelente capacidad del modelo para discriminar entre las clases de viviendas caras y no caras. Este alto valor de AUC asegura un buen equilibrio entre sensibilidad y especificidad, haciendo del modelo una herramienta fiable para evaluaciones de riesgo y decisiones de crédito hipotecario.

La eficiencia del modelo también se vio favorecida por la implementación de técnicas de regularización, que ayudaron a mitigar el sobreajuste a los datos de entrenamiento. Esto se evidenció en la consistencia de las métricas de rendimiento a través de los conjuntos de datos de entrenamiento y prueba, asegurando que el modelo generaliza bien a nuevos datos.

Referencias