Tecnológico de Monterrey

Analítica de datos y herramientas de inteligencia artificial II TI3001C

Reporte Actividad 7

Regresión logística

Integrantes del equipo:

Jerónimo Bernat Regordosa	A01735591
Estefanía López Ponce	A01654214
José Bryan Zamora Pacheco	A01707585
Manlio F. Rivera Pérez	A01734797

Octubre 19, 2023

Introducción

La regresión logística, es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre dos factores de datos (variables). Gracias a dicha relación, podemos predecir el valor de uno de esos factores basándose en el otro.

Algunos de los beneficios de utilizar esta técnica de análisis son:

- Simplicidad: Son matemáticamente menos complejos que otros métodos
- Velocidad: A pesar del volúmen de las bases de datos, esta técnica es la que requiere menos memoria y capacidad computacional.
- Flexibilidad: Se puede basar de acuerdo a 2 o más variables para los resultados o pronósticos.

A lo largo de este informe, construiremos múltiples modelos de regresión logística. Estos modelos se basan en las variables contenidas en la base de datos que hemos venido trabajando, la cual identifica a las personas que tienen incumplimientos de pago y sus respectivas características. El objetivo es utilizar estos modelos para predecir distintas variables relacionadas con este contexto.

Conversión de variables

El Data Frame en el que estamos trabajando contiene 12 variables, de las cuales 6 son de tipo categóricas y 6 son de tipo numéricas. Para desarrollar los modelos de regresión logística, es necesario que la variable dependiente, es decir, la variable que se desea predecir, sea de naturaleza binaria numérica. Al analizar este requisito, la única variable en el Data Frame que satisface estas características es "Risk_Flag". Esta variable se compone de 0 y 1, donde los 0 representan a las personas sin incumplimientos de pago en un préstamo, y los 1 representan a las personas con incumplimientos. Por lo tanto, procederemos a transformar todas las variables posibles en variables binarias de tipo numéricas, de modo que podamos emplearlas en nuestros modelos.

Comenzamos la transformación de las variables categóricas en variables binarias. El primer paso implicó analizar la cantidad de valores únicos contenidos en cada variable. Aquellas

variables que tenían un número limitado de valores únicos fueron seleccionadas para su conversión en variables binarias numéricas.

```
Variable: Married/Single
Contiene: 2 valores


Variable: House_Ownership
Contiene: 3 valores


Variable: Car_Ownership
Contiene: 2 valores


Variable: Profession
Contiene: 51 valores


Variable: CITY
Contiene: 317 valores


Variable: STATE
Contiene: 29 valores
```

Figura 1. Valores únicos en las variable de tipo categóricos del DataFrame

Observamos en la Figura 1 que las variables 'STATE', 'CITY', y 'Profession' no serán consideradas para la transformación, ya que contienen numerosos valores únicos que no pueden agruparse de manera efectiva en solo dos categorías. Por otro lado, las variables "Married/Single", "House_Orwnership" y "Car_Orwnership" sí serán convertidas en variables binarias numéricas, de la siguiente manera:

- Para la variable "Married/Single", el valor 0 representará "single" y el valor 1 representará "married".
- En cuanto a la variable "House_Orwnership", el valor 0 representará "owned", y el valor 1 representará los valores "rented" y "nonorent_noown".
- Por último, para la variable "Car_Orwnership", el valor 0 representará "no", y el valor 1 representará "yes".

Luego continuamos con las variables numéricas. En este tipo de variable, calculamos la media de los datos y clasificamos los valores por debajo de la media como 0 y los valores por encima de la media como 1. Esta es una forma simple y rápida de categorización, pero consideramos que es una adecuada manera de dividir los datos para su uso en modelos de regresión logística.

Modelos de regresión logística

Desarrollamos un total de 10 modelos de regresión logística, en cada uno seleccionamos variables dependientes e independientes diferentes, en función de la predicción que se quería realizar. En todos los escenarios, la variable dependiente tuvo una naturaleza binaria numérica y las variables independientes una naturaleza numérica.

A continuación, se presenta los resultados de los modelos desarrollados:

Tabla 1. Comparación de las métricas obtenidas de las regresiones logísticas realizadas.

Variable dependiente (y)	**Variables independiente (X)**	**Clase objetivo**	**Precisión %**	**Exactitud %**	**Sensibilidad %**
Risk_Flag	Age, Income, Experience	0	0.877	0.877	1.0
Income_binary	Age, Experience	0	0.499	0.502	0.556
Income_binary	Experience, CURRENT_JOB_YRS	0	0.4997	0.4994	0.2893
Experience_ binary	Age, Income , CURRENT_JOB_YRS	0	0.710	0.713	0.757
Income_binary	Age , Car_Ownership_binary, House_Ownership_binary, Experience	0	0.506	0.508	0.529
Age_binary	Experience, CURRENT_HOUSE_YRS , House_Ownership_binary, Married/Single_binary	0	0.511	0.510	0.757
Age_binary	CURRENT_JOB_YRS, CURRENT_HOUSE_YRS	0	0.508	0.510	0.208
Income_binary	Age_binary, Experience	1	0.502	0.502	0.496
CURRENT_JOB_ YRS binary	Income,Experience	0	0.720	0.694	0.736
Experience_binary	House_Ownership_binary', ‘CURRENT_JOB_YRS_bi nary’	0	0.708	0.710	0.754

Hallazgos

Modelo 1

- Variable dependiente: Risk_Flag
- Variables independientes: Age, Income, Experience

En la Tabla 1, observamos que el primer modelo exhibe un nivel elevado de precisión y exactitud al intentar predecir la clase 0, con valores de ambas métricas en 0.877. Además, la sensibilidad es del 100%, indicando que el modelo no omite ningún valor de 0, eliminando así los falsos negativos. Estos resultados sugieren un rendimiento excepcional de este modelo en la predicción de la clase 0 de la variable. Sin embargo, una revisión de la matriz de confusión revela que el modelo etiquetó todas las instancias como clase 0. Por lo tanto, podemos concluir que este modelo carece de utilidad, ya que no es capaz de predecir las clases en cuestión.

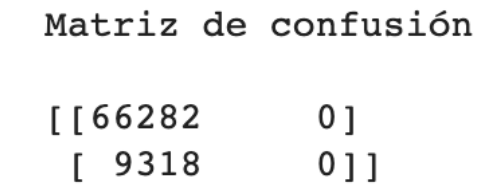

Figura 2. Matriz de confusión del modelo 1

Modelo 4

- Variable dependiente: Experience_binary
- Variables independientes: Age, Income, CURRENT_JOB_YRS

En conclusión, descartando el primer modelo, de los modelos restantes, el modelo más destacado es el número 4, diseñado para predecir la clase 0 en la variable "Experience_binary". Este modelo presenta los porcentajes más altos de exactitud y sensibilidad en comparación con los demás modelos. Esto sugiere que el modelo es eficaz en la predicción de la experiencia laboral por debajo del promedio. Con una sensibilidad del 75.7%, el modelo demuestra un rendimiento sólido al identificar correctamente la mayoría de las instancias de la clase 0, minimizando los falsos negativos. Es importante destacar que, en las tres métricas evaluadas, este modelo supera el umbral de 0.7, lo que respalda su eficacia en la predicción de la clase 0 de la variable dependiente.