Actividad 4.2: Regresión Logística

Sebastián Mejía A01722536

Juan Zamorano A01642396

Carlos Fonseca A01734538

Emily Bueno A01736939

Profesor: Alfredo Garcia Suarez

Tecnológico de Monterrey

Campus Puebla

26/04/2025

Introducción

En esta actividad trabajamos con una base de datos que contenía varias características sobre el comportamiento de usuarios en un sistema o plataforma. El objetivo principal fue utilizar la técnica de **regresión logística** para entender mejor cómo unas variables pueden ayudarnos a predecir otras. La regresión logística se usa cuando el resultado que queremos predecir solo tiene dos opciones posibles, como por ejemplo si una persona acepta o no una oferta.

Carga y Preparación de los Datos

Lo primero que se hizo fue cargar la base de datos en el entorno de trabajo. Se revisó que los datos estuvieran completos y que las columnas fueran entendibles. Después, se realizó una transformación de variables. Esto significa que algunas columnas que originalmente tenían varios valores posibles se convirtieron en variables más sencillas, que solo pueden tomar dos valores (por ejemplo, 0 o 1). Esta transformación fue muy importante porque la regresión logística solo funciona bien si las variables que analizamos tienen este tipo de formato sencillo.

Selección de Casos

Luego, se seleccionaron diferentes combinaciones de variables para hacer el análisis. Se trabajaron diez casos diferentes, en los cuales se eligieron una variable dependiente (lo que se quiere predecir) y una o más variables independientes (lo que se utiliza para hacer la predicción). En cada uno de los casos, se preparó el conjunto de datos dividiéndolo en dos partes: una para entrenar el modelo y otra para probarlo. Esta división nos permitió verificar si el modelo realmente aprendía a predecir o si simplemente se estaba "memorizando" los datos.

Creación y Evaluación de los Modelos

Después de entrenar los modelos, se analizaron los **resultados**. Para cada caso, se imprimieron los **coeficientes** (que indican qué tanto influye cada variable en la

predicción) y se generaron **matrices de confusión** que ayudan a ver cuántas predicciones fueron correctas y cuántas no. También se calculó la **precisión** de cada modelo, que básicamente nos dice qué porcentaje de predicciones fueron correctas en el conjunto de prueba.

Durante el proceso también se realizaron algunas **mejoras al código**. Se eliminaron configuraciones que no eran necesarias, como el parámetro zero_division, haciendo que el código quedara más sencillo, más ordenado y fácil de leer. Aunque este parámetro ayuda a evitar errores cuando se presentan divisiones por cero, en este caso no era necesario porque los datos ya estaban preparados correctamente.

Principales Hallazgos

Al analizar los resultados obtenidos, uno de los hallazgos más importantes fue que **la mayoría de los modelos funcionaron muy bien**. En varios de los casos, se obtuvo una **precisión superior al 85%**, lo cual indica que las variables elegidas eran buenas predictores de los resultados. Esto quiere decir que, basándose en las variables que seleccionamos, el modelo podía predecir correctamente la mayoría de las veces si algo iba a pasar o no.

Otro punto importante que se observó fue que **la preparación de los datos** fue clave para el análisis. Si las variables no se hubieran transformado correctamente en variables binariaso dicotomicas, los modelos no habrían funcionado o habrían dado resultados incorrectos. Así que preparar los datos correctamente no solo es un paso obligatorio, sino que también es esencial para obtener buenos resultados.

Aunque la mayoría de los modelos mostraron resultados muy buenos, también hubo algunos casos donde la precisión fue más baja, cerca del 70%. Estos casos nos indican que quizá se necesitan buscar otras variables independientes que estén más relacionadas con el resultado o tal vez trabajar más en la transformación de los datos para que el modelo pueda entender mejor las relaciones. No siempre es posible lograr una precisión perfecta, pero reconocer cuándo un modelo no es tan bueno también es parte importante del análisis.

Una cosa importante de mencionar, es que hubo una parte del codigo la cual no pudo ser procesada, en el momento de realizar los modelos para los usuarios

“ADRIAN”, “ALEIDA” , “ARLETT”, “ASHLEY y “AUSTIN”; no se pudieron procesar debido a que habia falta de datos o algunas columnas se eliminaron por la misma situacion, se siguieron todos los pasos correspondientes, sin embargo, no se podia realizar debido a esto.

Finalmente, el trabajo demuestra que, utilizando técnicas como la regresión logística, podemos construir modelos predictivos útiles a partir de bases de datos. Además, muestra que la organización del código y la limpieza de los datos son pasos fundamentales para que todo el proceso funcione de manera correcta.

Conclusión

En conclusión, esta actividad permitió entender de manera práctica cómo aplicar la regresión logística en bases de datos reales, demostrando la importancia de preparar adecuadamente los datos, interpretar los resultados de forma objetiva y buscar siempre mejorar los modelos cuando sus resultados no son los esperados.