Tecnológico de Monterrey

Modelos de Clasificación Lineal

–

Minería de Datos

Mónica González - A01735626
Mariana Rico - A01735770
Alberto Muro - A01734046

Base de Datos

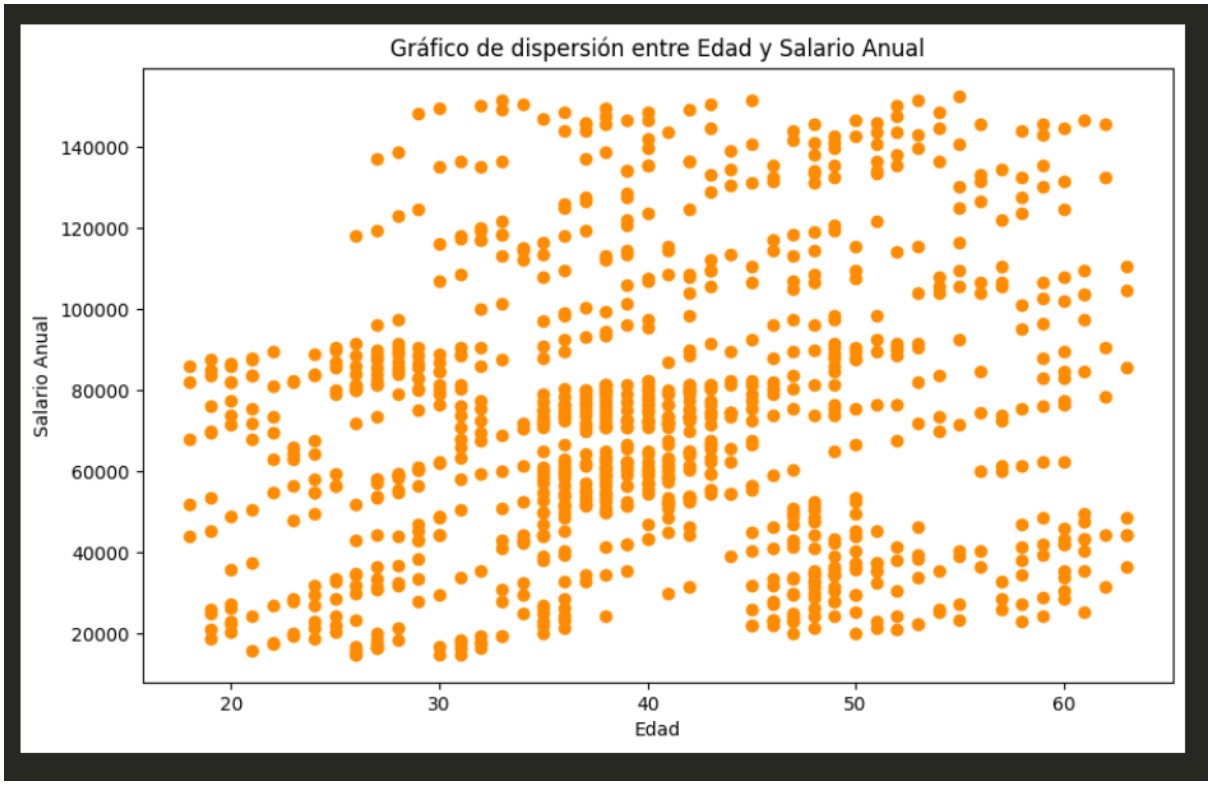
En este ejercicio, trabajaremos con la base de datos llamada "car_data," que incluye diversas variables relacionadas con la adquisición de vehículos, como género, edad, salario anual y el hecho de si se ha producido una adquisición o no. El objetivo de este estudio es aplicar dos tipos de modelos lineales: el Análisis Discriminante Lineal y la Regresión Logística. Utilizaremos **adquisición (purchased)** como nuestra variable dependiente y **edad** y **salario anual** como variables independientes.

Para evaluar la eficacia de nuestros modelos, dividiremos la base de datos en dos conjuntos: uno para el entrenamiento de los modelos y otro para su prueba y validación. Esto nos permitirá estimar el rendimiento del modelo en datos no vistos. Este ejercicio de modelado de datos tiene como objetivo utilizar dos enfoques diferentes para abordar el problema de la adquisición de vehículos a partir de variables como la edad y el salario anual. La elección entre el Análisis Discriminante Lineal y la Regresión Logística dependerá de la calidad de los modelos y su capacidad para predecir con precisión la adquisición de vehículos en función de los datos proporcionados en la base de datos "car_data."

Análisis discriminante lineal.

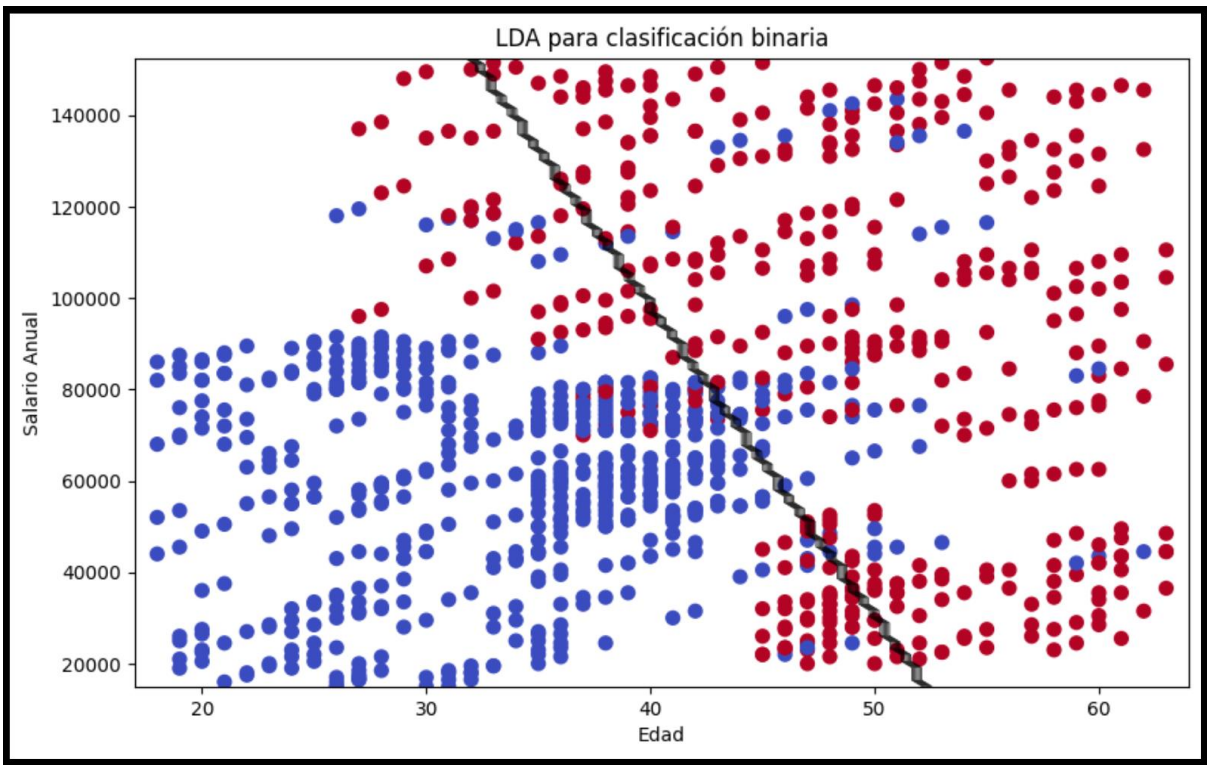
Interacción de variables independientes.

Antes de evaluar el modelo comparamos mediante un gráfico de dispersión la interacción de nuestras variables independientes, en este observamos que a simple vista no existe ningún tipo de relación entre estas, lo cual podría darnos un modelo de clasificación lineal no tan bueno.



Interacción de variables independientes categorizadas.

Según nuestro LDA observamos una clara distinción entre las personas que adquieren un vehículo y las que no siendo las primeras representadas de color rojo dentro del gráfico, lo cual nos habla de un perfil donde una mayor edad y un salario anual alto influyen en la adquisición del vehículo. Por otro lado, la categoría de color azul representa al grupo de personas que no adquieren el vehículo, los cuales presentan una edad mucho mas joven con un menor salario anual.



Evaluación del Modelo

Al correr el modelo observamos que obtenemos buenos resultados de este, como se muestra a continuación:

	precision	recall	f1-score	support
No Adquirido	0.83	0.90	0.86	598
Adquirido	0.83	0.72	0.77	402
accuracy			0.83	1000
macro avg	0.83	0.81	0.82	1000
weighted avg	0.83	0.83	0.82	1000

Nos percatamos que el nivel de precisión es de 0.83 en ambas categorías, lo cual nos habla del nivel de proporción de las muestras clasificadas correctamente en comparación con el total de las muestras (al igual que en exactitud).

En la sensibilidad medimos la capacidad del modelo para identificar todas las muestras positivas apropiadamente (o tasa de verdaderos positivos), donde tenemos una mejor identificación de estos para los que no adquieren vehículo con 0.90 y con 0.72 para los que, si lo adquieren, finalmente dentro de mi valor f-1 (que es una ponderación entre sensibilidad y precisión) cuento con un valor de 0.86 en la categoría “No Adquirido” y 0.77 para “Adquirido”.
Aparentemente mi modelo podría ser “bueno” considerando los resultados obtenidos, para comprobar esto haremos uso de las siguientes metodologías.

Validación Cruzada

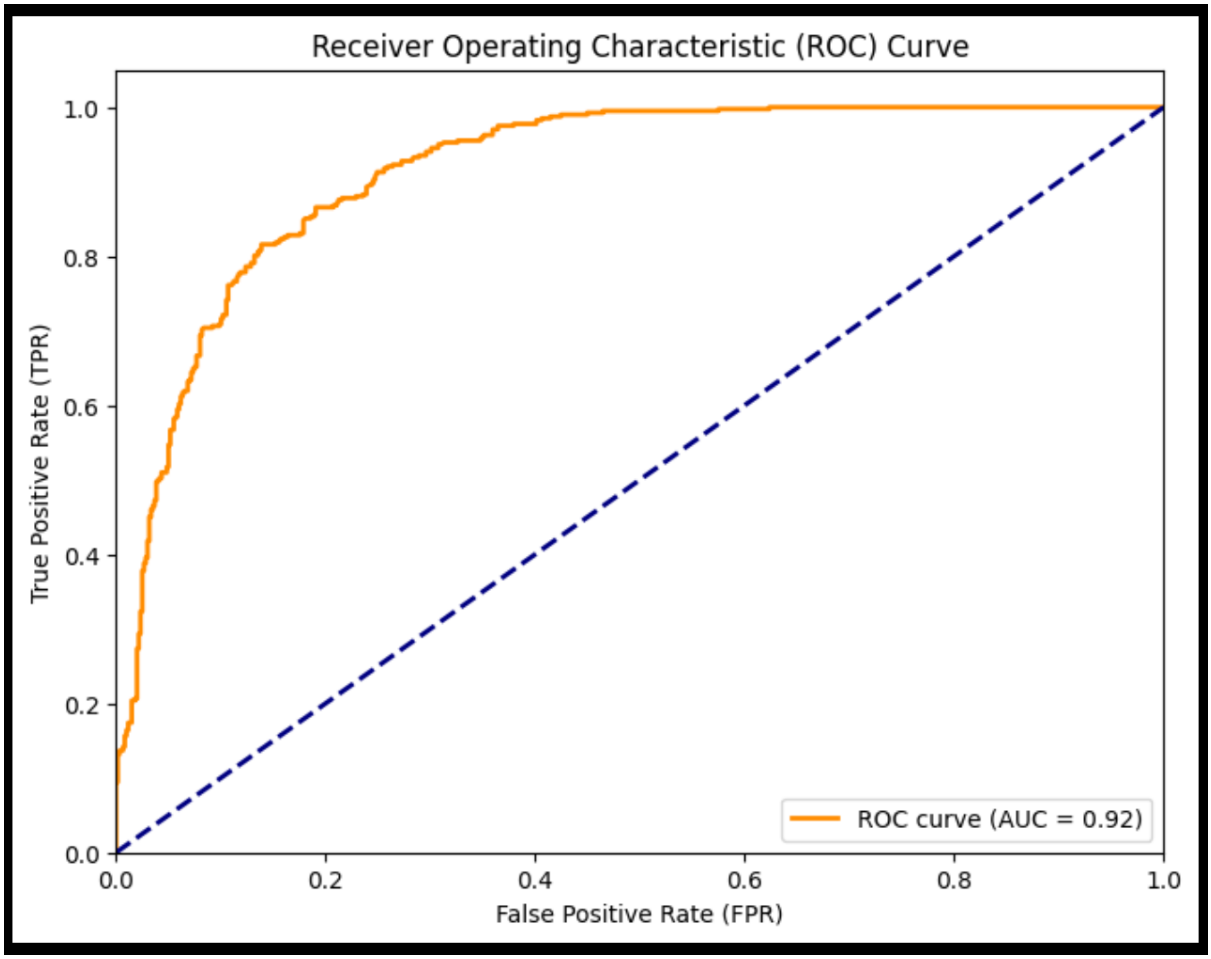
Al utilizar validación cruzada estamos entrenando el modelo dividimos la base entre el conjunto de datos de entrenamiento y el de prueba, de esta manera corremos este un total de 30 veces para determinar la precisión de este, finalmente obtenemos un promedio de cada uno de los resultados y obtenemos una precisión promedio de 0.83 con una desviación de 0.0690.

```
Puntaje de validación cruzada: [0.85294118 0.73529412 0.94117647 0.85294118 0.70588235 0.73529412
 0.82352941 0.88235294 0.88235294 0.73529412 0.81818182 0.81818182
 0.96969697 0.87878788 0.93939394 0.90909091 0.78787879 0.81818182
 0.90909091 0.75757576 0.75757576 0.84848485 0.84848485 0.84848485
 0.75757576 0.84848485 0.75757576 0.84848485 0.81818182 0.72727273]
Precisión promedio: 0.8271241830065359
Desviación estándar de la precisión: 0.06902888330801567
```

ROC y AUC

Finalmente, exponemos de manera grafica la sensibilidad del modelo mediante la curva ROC, donde en el valor AUC obtenemos una puntuación de 0.92, esto nos habla sobre la capacidad que tiene el modelo para discriminar entre las clases positivas y las negativas de los datos.

Al ser un valor muy cercano a 1 declaramos que el modelo tiene una gran capacidad para hacer discriminaciones entre los conjuntos de datos de una manera muy buena.



Regresión Logística Binaria

Definición de Variables

Al igual que en el modelo LDA utilizaremos “edad” y “salario anual” como independientes y “adquisición” como dependiente.

Evaluación del Modelo

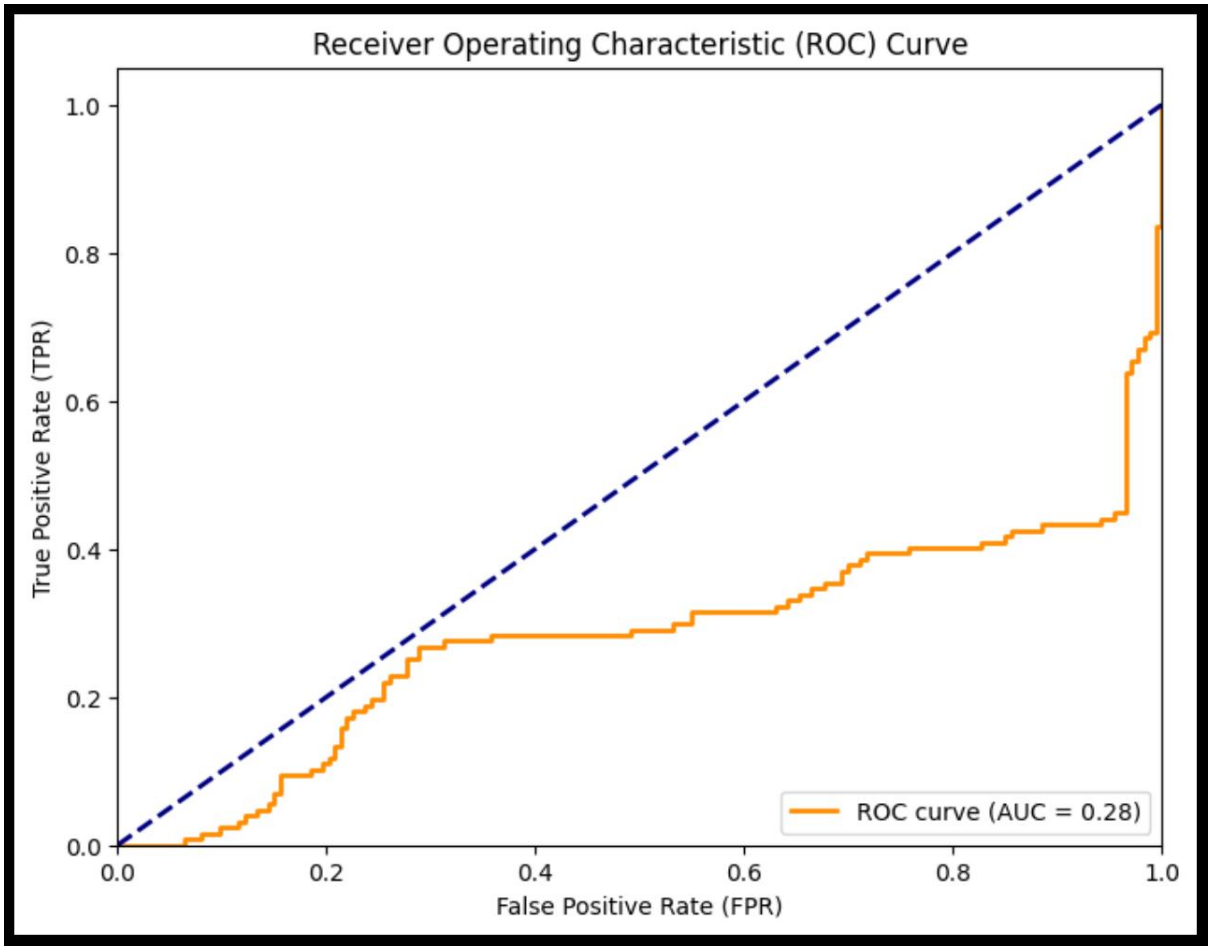
En este modelo siempre obtenemos un reporte de clasificación con números diferentes, se hicieron distintas pruebas alterando la proporción de entrenamiento y prueba del conjunto de datos y aun así el resultado más común es el que se muestra a continuación.

```
Confusion Matrix:
[[173   0]
 [127   0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.58      1.00      0.73       173
           1       0.00      0.00      0.00       127

    accuracy                           0.58       300
   macro avg       0.29      0.50      0.37       300
weighted avg       0.33      0.58      0.42       300
```

El modelo no logra hacer una correcta discriminación entre las categorías, tiene los mejores resultados al analizar “No Adquirido” y aun así las proporciones de esta oscilan bastante en comparación con las demás métricas.

ROC y AUC

Finalmente, en nuestro gráfico observamos que nuestra área bajo la curva (AUC) tiene una puntuación de 0.28, lo cual lo hace un modelo inútil para realizar una correcta discriminación, de esta forma que lograríamos tener mejores resultados si ponemos en juego el modelo con el factor de aleatoriedad, de esta manera obtendríamos un valor de 0.5.



Modelo Final

Con la información obtenida con el programa, obtenemos la siguiente ecuación como nuestro modelo matemático:

$$logit = Li = -1.27587836e-10 \ - 1.06897495e-09X_1 \ - 1.17503473e-06X_2$$

Sin embargo, esta no es confiable debido a los resultados de la evaluación del modelo obtenidos anteriormente.

Conclusión

En el presente análisis observamos que no existe una relación evidente entre las variables independientes "edad" y "salario anual". Esto sugiere un posible desafío en la implementación de un modelo de clasificación lineal efectivo. Sin embargo, el análisis discriminante lineal (LDA) reveló una clara distinción entre las personas que adquieren un vehículo y las que no. Las personas que adquieren un vehículo tienden a ser de mayor edad y tener un salario anual más alto, mientras que las que no adquieren un vehículo son más jóvenes y tienen ingresos más bajos.

A diferencia del LDA, la Regresión Logística Binaria mostró dificultades para discriminar entre las categorías. Los resultados variaron considerablemente al modificar la proporción de datos de entrenamiento y prueba, y el modelo tuvo dificultades para clasificar de manera consistente las categorías "Adquisición" y "No Adquisición."

En definitiva, el modelo de Análisis Discriminante Lineal demostró ser más efectivo en este ejercicio, con una buena capacidad para predecir la adquisición de vehículos basándose en la edad y el salario anual de los individuos. La Regresión Logística Binaria, por otro lado, mostró dificultades en la discriminación de las categorías. Esto sugiere que el LDA es la mejor opción para predecir la adquisición de vehículos.