Introducción

El objetivo de este proyecto es desarrollar un modelo de predicción de preferencias de compra basado en variables sociodemográficas de los clientes, como el salario estimado y la edad. Para ello, se utilizan técnicas de análisis y modelado de datos, evaluando la capacidad predictiva del modelo mediante diversas pruebas. El trabajo se realizó utilizando los lenguajes de programación Python y R.

Librerías Utilizadas En Python:

- pandas: Permite la manipulación y análisis de datos a través de estructuras como DataFrames y Series.
- seaborn: Facilita la visualización de datos estadísticos con gráficos informativos y estilizados.
- numpy: Proporciona soporte para operaciones matemáticas y manipulación eficiente de matrices y arreglos numéricos.
- statsmodels.api: Ofrece herramientas avanzadas de modelado estadístico, incluyendo regresión y pruebas de hipótesis.
- sklearn.preprocessing.StandardScaler: Estandariza los datos para garantizar que todas las variables tengan la misma escala en el modelo.
- sklearn.metrics: Permite calcular métricas de evaluación de modelos, como la curva ROC, matriz de confusión, accuracy score y AUC-ROC.
- matplotlib.pyplot: Brinda funcionalidades para crear visualizaciones como histogramas, gráficos de dispersión y curvas ROC.

En R:

- readxl: Permite leer archivos de Excel para importar la base de datos.
- dplyr: Facilita la manipulación y transformación de datos mediante funciones intuitivas.
- ggplot2: Proporciona herramientas para la creación de visualizaciones estéticas y altamente personalizables.
- pROC: Se utiliza para calcular y visualizar curvas ROC, que ayudan a evaluar la discriminación del modelo.
- caret: Proporciona herramientas para la preparación de datos, selección de modelos y evaluación de desempeño.
- broom: Convierte objetos estadísticos en tibbles para facilitar su manipulación y visualización.
- ineq: Permite calcular índices de desigualdad, como el índice de Gini, usado en la validación del modelo.

Desarrollo del modelo

Como primer paso en este proyecto, importamos las librerías que usaremos a lo largo del desarrollo del modelo, junto con la carga de la base de datos que se utilizará y la especificación del nombre de la hoja que usaremos, en este caso "E11".

Se realiza una segmentación de los datos en conjuntos de entrenamiento y prueba. Esta división se efectúa en función del valor de la columna "Sample", categorizada en "Train" y "Test".

Para generar el modelo predictivo, se llevan a cabo los siguientes pasos:

- Conversión de variables categóricas a valores numéricos.
- Estandarización de los datos para garantizar una escala uniforme en las variables.

Se construye un primer modelo utilizando las variables "EstimatedSalary USD", "Age" y "Gender". Sin embargo, tras aplicar la prueba Z, se identifica que la variable "Gender" no cumple con los criterios estadísticos requeridos, lo que motiva su eliminación del modelo.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:              Purchased   No. Observations:                  323
Model:                          Logit   Df Residuals:                      319
Method:                           MLE   Df Model:                            3
Date:                Sun, 16 Feb 2025   Pseudo R-squ.:                  0.4671
Time:                        17:32:44   Log-Likelihood:                -111.76
converged:                       True   LL-Null:                       -209.71
Covariance Type:            nonrobust   LLR p-value:                 3.237e-42
===========================================================================================
                              coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                   -1.1549      0.189     -6.124      0.000      -1.525      -0.785
EstimatedSalary USD      1.1197      0.204      5.489      0.000       0.720       1.519
Gender                   0.1974      0.170      1.160      0.246      -0.136       0.531
Age                      2.4925      0.305      8.184      0.000       1.896       3.090
===========================================================================================
```

Se genera un segundo modelo sin la variable "Gender". Al someterlo nuevamente a la prueba Z, se verifica que cumple con los criterios establecidos, lo que indica su viabilidad.

Logit Regression Results

Dep. Variable:	Purchased	No. Observations:	323
Model:	Logit	Df Residuals:	319
Method:	MLE	Df Model:	3
Date:	Sun, 16 Feb 2025	Pseudo R-squ.:	0.4671
Time:	17:32:44	Log-Likelihood:	-111.76
converged:	True	LL-Null:	-209.71
Covariance Type:	nonrobust	LLR p-value:	3.237e-42

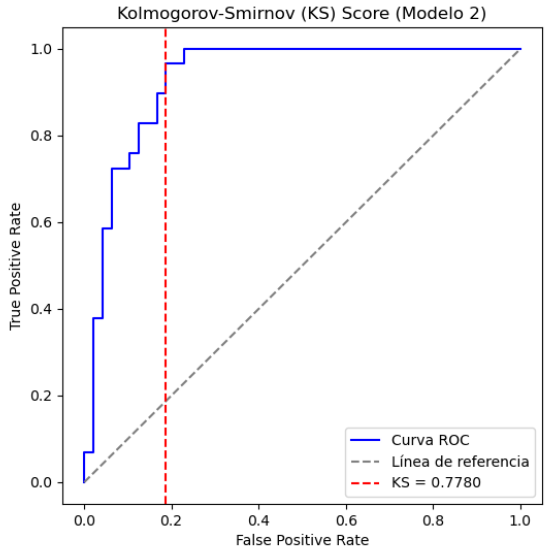
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1549	0.189	-6.124	0.000	-1.525	-0.785
EstimatedSalary USD	1.1197	0.204	5.489	0.000	0.720	1.519
Gender	0.1974	0.170	1.160	0.246	-0.136	0.531
Age	2.4925	0.305	8.184	0.000	1.896	3.090

Para evaluar la capacidad predictiva y discriminativa del modelo, se aplican las siguientes pruebas:

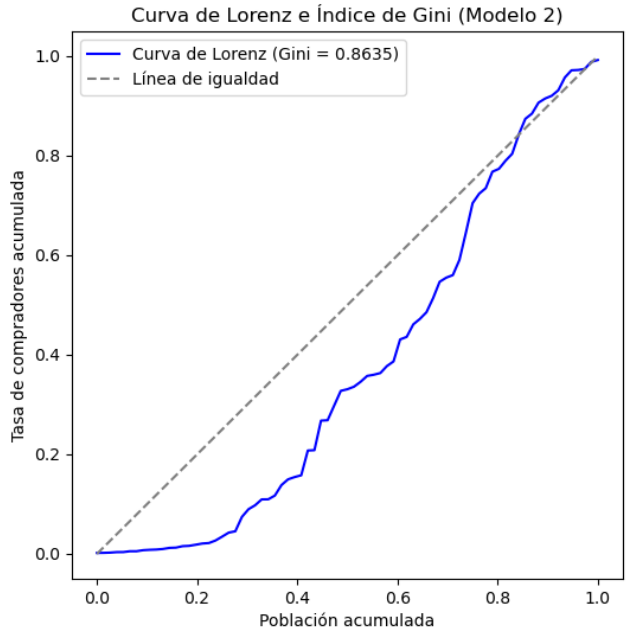
- Curva ROC
- Prueba KS (Kolmogorov-Smirnov)
- Índice de Gini

Los resultados obtenidos en estas pruebas fueron:

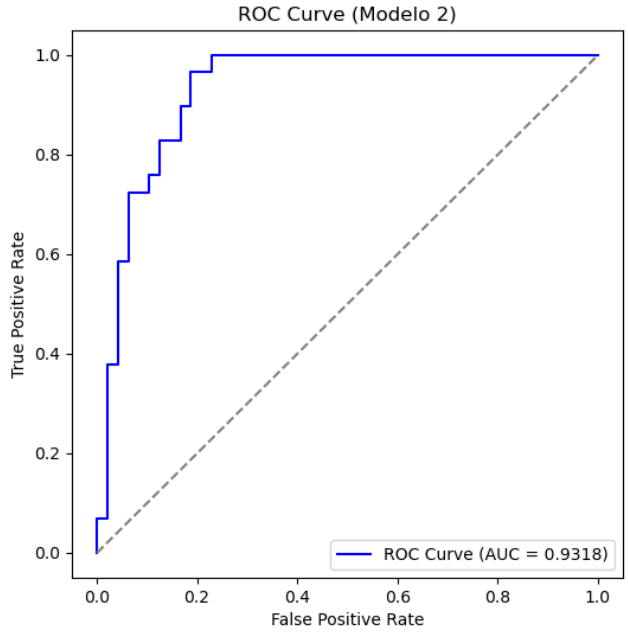
- KS Score: 0.7780



- Gini Index: 0.8635

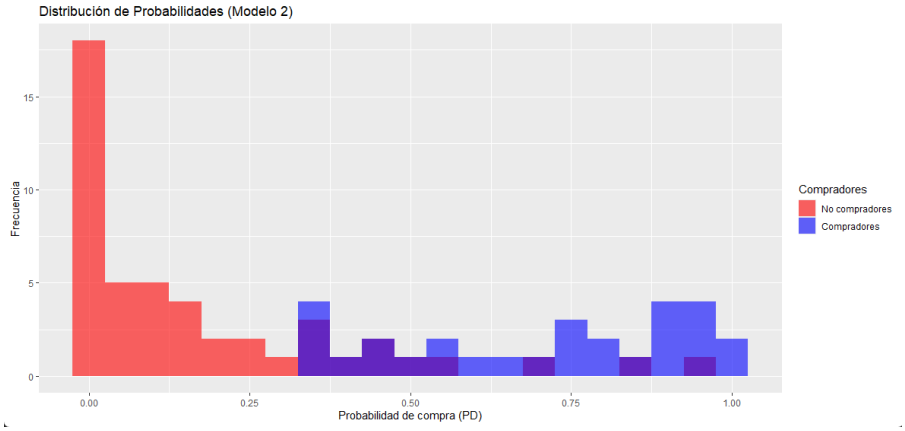


- AUC Score: 0.9318

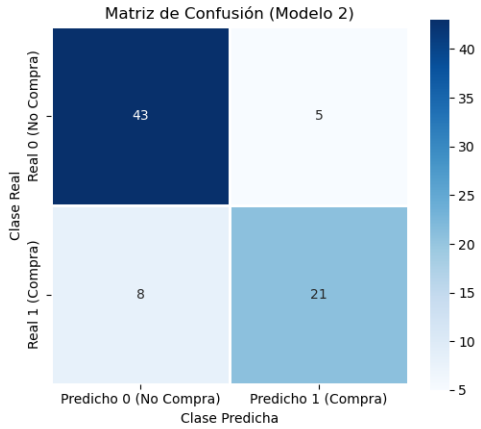


Dado que las tres pruebas arrojaron resultados superiores a 0.75, podemos concluir que el modelo tiene una capacidad predictiva y discriminativa alta, haciéndolo un modelo confiable para este tipo de análisis.

Una vez realizado lo anterior, revisamos la precisión de nuestro modelo usando los datos de prueba para revisar la precisión del modelo creado con los datos de entrenamiento, en esta gráfica podemos ver la frecuencia en que se distribuyen los compradores y los no compradores según nuestro modelo.



Una vez visto lo anterior, realizamos una matriz de confusión para conocer los aciertos y errores de nuestro modelo. Así podemos ver que nuestro modelo tiene una precisión aproximada del 83.11%.



Conclusiones

Con base en el análisis realizado, se concluye que las variables "EstimatedSalary USD" y "Age" son determinantes en la predicción de preferencias de compra. El modelo desarrollado presenta una eficiencia superior al 80%, lo que lo convierte en una herramienta valiosa para el análisis y segmentación de clientes en entornos comerciales.