Universidad Tecnológica Nacional FRRo

Minería de datos

Comisión 5E05

Trabajo Práctico Integrador

Etapa 1

Grupo Nº 5

Integrantes:

47066 - Gorosito, Adriel

47447 - Botali, Santiago

29694 - Jimenez, Dana Marina

27836 - Torres, Paula

Índice

Contexto

Problema

El gerente de una empresa que se dedica a la venta de productos para el hogar desea mantener un buen posicionamiento en el mercado. Para ello, elaboró una serie de estrategias comerciales para el siguiente año.
Una de las estrategias planteadas es aumentar la venta de bicicletas gracias a un convenio con otra empresa. Se fabricarán tres tipos de bicicletas:

- Bicicletas para niños (Kinder)
- Bicicletas estándares (Basic)
- Bicicletas deportivas (Sport)

Para poder llevar a cabo esto, el sector de marketing necesita ayuda en la campaña publicitaria por correo electrónico. Para ello, se posee un archivo de 1500 potenciales clientes, sobre los cuales hay que decidir si se le envía o no la publicidad y el contenido del correo.

Objetivos

- Establecer un criterio de elección de cliente potencial.
- Decidir a qué clientes mandarle la publicidad, teniendo en cuenta que es mejor mandarle la publicidad a alguien desinteresado que tener una pérdida de una venta.
- Clasificar a los clientes para determinar el tipo de bicicleta que le podría llegar a interesar, para luego poder realizar marketing personalizado.

Análisis exploratorio

Introducción

Con los datos que disponemos, a continuación realizaremos un análisis exploratorio para obtener información necesaria para el estudio de la información, utilizando nuestro criterio para reconocer aquellos datos que no son relevantes y aquellos que sí.

Comenzamos generando una tabla que nos permita ver, de forma general, todas las variables. Esta tabla nos muestra, para cada variable, el nombre, la cantidad de entradas no nulas y el tipo de dato. Además, informa de manera general, el rango de los datos (el cual es de 6400) y la cantidad de tipos de datos.

```
RangeIndex: 6400 entries, 0 to 6399
Data columns (total 21 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   IdCliente           6400 non-null   int64
 1   IdCiudad            6400 non-null   int64
 2   Nombre              6400 non-null   object
 3   Apellido            6400 non-null   object
 4   FechaNacimiento     6400 non-null   object
 5   EstadoCivil         6400 non-null   object
 6   Genero              6400 non-null   object
 7   Email               6400 non-null   object
 8   IngresoAnual        6390 non-null   float64
 9   TotalHijos          6400 non-null   int64
 10  Educacion           6400 non-null   object
 11  Ocupacion           6400 non-null   object
 12  Propietario         6400 non-null   int64
 13  CantAutomoviles     6400 non-null   int64
 14  Direccion           6400 non-null   object
 15  Telefono            6400 non-null   object
 16  FechaPrimeraCompra  6400 non-null   object
 17  Distancia           6400 non-null   object
 18  Region              6400 non-null   object
 19  Edad                6400 non-null   int64
 20  ComproBicicleta     6400 non-null   int64
dtypes: float64(1), int64(7), object(13)
```

Tabla 1: Primera vista general de los datos.

Posteriormente, generamos una matriz donde las columnas son las variables de interés y las filas son estadísticos.

	IngresoAnual	TotalHijos	Propietario	CantAutomoviles	Edad	ComproBicicleta
count	6390.000000	6400.000000	6400.000000	6400.000000	6400.000000	6400.000000
mean	57532.081377	1.894844	0.676562	1.547656	51.195469	0.394375
std	32331.969091	1.630993	0.467825	1.147060	11.517698	0.488754
min	10000.000000	0.000000	0.000000	0.000000	32.000000	0.000000
max	170000.000000	5.000000	1.000000	4.000000	102.000000	1.000000

Matriz 1: Estadísticos para las variables de interés.

En las filas, **count** es el total de datos *no nulos* analizados, **mean** es la media de los datos, **std** es el desvío estándar (distancia promedio de los datos a la media), **min** es el valor mínimo de la columna (por ejemplo, el ingreso anual mínimo percibido es de $10000) y **max** el valor máximo de la misma (el ingreso anual máximo percibido es de $170000).

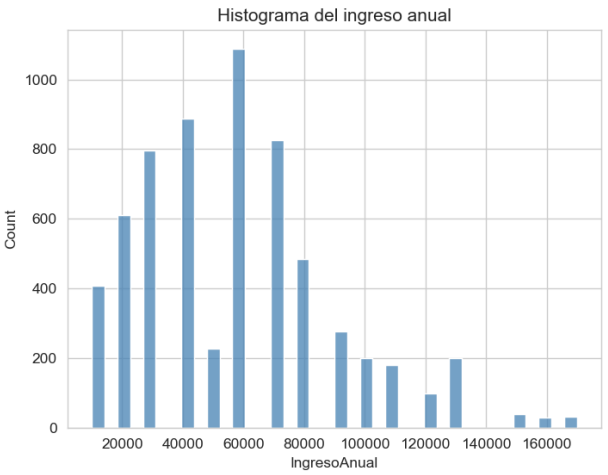
Esto nos brinda una información general de los clientes ya percibidos, que nos servirá para proceder con nuestro análisis.

Valores nulos

El primer desperfecto que encontramos, es que en **IngresoAnual** hay 6390 datos no nulos (como el rango es de 6400, entonces hay 10 valores nulos). Para este caso, hay dos formas de proceder: dejar los valores nulos o rellenarlos. Nosotros elegimos rellenarlos debido a que deseamos que todas las variables tengan la misma cantidad de datos.

Rellenado de valores nulos

Para rellenar estos valores, podemos utilizar la media o la mediana. En general, se recomienda utilizar la media si los datos siguen una distribución normal. Para averiguarlo, generamos un histograma para analizar la distribución de los datos.



Gráfica 1: Histograma del ingreso anual.

Como podemos ver, la distribución no es normal, sino que es sesgada a la derecha. Luego, decidimos imputar los valores nulos con la mediana.

En las siguientes tablas se pueden ver el antes y después de la imputación:

	IngresoAnual	IngresoAnual
count	6390.000000	6400.000000
mean	57532.081377	57535.937500
std	32331.969091	32306.842992
min	10000.000000	10000.000000
max	170000.000000	170000.000000

Tabla 2: Estadísticos del ingreso anual (pre-rellenado y post-rellenado, respectivamente).

Observamos que varía sólo la media y el desvío; el mínimo, máximo y los cuartiles siguen siendo iguales (esto es así ya que se rellena con la mediana). No consideramos que el hecho de imputar los datos haya sido factible debido a la casi insignificante variabilidad que ocasionó, pero si lo consideramos necesario para obtener un resultado mas especifico y completo del estudio.

Variable Objetivo

Analizaremos la columna **ComproBicicleta**. Esta es la variable que interesa estudiar, ya que parecería informar si el cliente compró una bicicleta anteriormente o no. Observando la media e interpretando, el 39% de las personas compraron bicicletas.
Sin embargo, el desvío es mayor que la media, sugiriendo que hay una gran variabilidad en las decisiones de compra de bicicletas entre los clientes. Esto significa que hay muchos factores diferentes que pueden influir en la decisión de comprar una bicicleta, pudiéndo ser algunos de estos factores la edad, el ingreso anual, si tiene auto o no, etc. Por lo tanto,

nuestro objetivo es analizar estos factores, intentando buscar alguna relación para así poder generar una conclusión.

Volviendo a la variable objetivo, generando algunos de sus valores nos encontramos que se trata de una columna booleana (es decir, toma solo dos valores, 0 si es falso y 1 si es verdadero).

	ComproBicicleta
0	3876
1	2524

Tabla 3: Valores que adquiere la variable ComproBicicleta.

Por lo tanto, según la tabla, 3876 personas no compraron bicicleta en el pasado y 2524 sí lo hicieron.

Correlación entre las variables

Matriz S

Generamos una matriz S para ver la correlación entre las variables:

	IngresoAnual	TotalHijos	Propietario	CantAutomoviles	Edad	ComproBicicleta
IdCliente	-0.048255	0.002471	-0.125417	0.026091	-0.010823	0.007200
IdCiudad	0.043773	0.061364	0.065481	-0.078792	0.080310	-0.085684
IngresoAnual	1.000000	0.222296	0.044658	0.469289	0.153101	0.054093
TotalHijos	0.222296	1.000000	0.185421	0.272527	0.495425	-0.131266
Propietario	0.044658	0.185421	1.000000	-0.054269	0.112114	0.020064
CantAutomoviles	0.469289	0.272527	-0.054269	1.000000	0.169977	-0.183216
Edad	0.153101	0.495425	0.112114	0.169977	1.000000	-0.101642
ComproBicicleta	0.054093	-0.131266	0.020064	-0.183216	-0.101642	1.000000

Matriz 2: Matriz S.

En la matriz S, un valor cercano a 1 o -1 indica una gran relación entre el par de variables.

- Si el valor es cercano a 1, significa que si una variable aumenta, la otra tiende a aumentar.
- Si el valor es cercano a -1, significa que si una variable aumenta, la otra tiende a disminuir.

Además, si el valor absoluto del valor obtenido es mayor o igual a 0.75, entonces se considera que las variables se correlacionan entre sí.

Sin embargo, creemos que solo interesa ver la correlación con **ComproBicicleta,** ya que la misma nos permite identificar aquellos individuos que ya han mostrado interés en las bicicletas y que podrían ser potenciales clientes. Para ello, generamos otra matriz S:

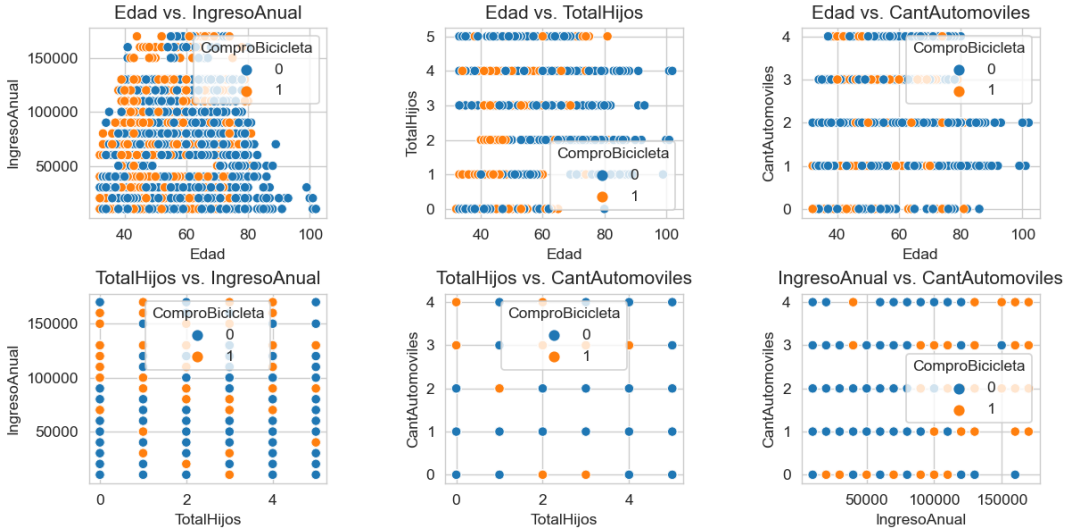
	IngresoAnual	TotalHijos	Propietario	CantAutomoviles	Edad	ComproBicicleta
ComproBicicleta	0.054093	-0.131266	0.020064	-0.183216	-0.101642	1.000000

En la tabla se puede ver que no hay valores cercanos a 1 ni a -1, por lo tanto podemos decir que no hay correlación directa entre el hecho de haber comprado una bicicleta y las demás variables. La variable que más se acerca es **CantAutomoviles**, con -0.18. Esto significa que, a medida que aumenta el número de automóviles de una persona, es menos probable

que hayan comprado una bicicleta en el pasado (esto podría deberse a la preferencia de utilizar un auto en lugar de una bicicleta para el transporte). Sin embargo, su correlación es demasiado chica como para ser tenida en cuenta, por lo que no es relevante.

Diagramas de dispersión estratificados

En conjunto a la matriz S, generamos algunos diagramas de dispersión estratificados entre las demás variables:

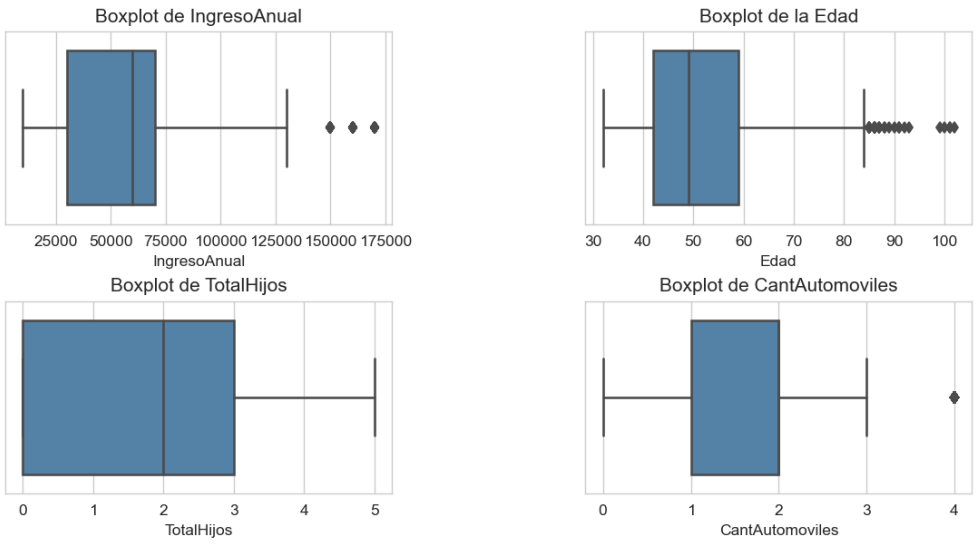


Gráfica 2: Diagramas de dispersión estratificados entre las demás variables.

Analizando todos los gráficos obtenidos, llegamos a la conclusión de que ningún diagrama aporta información relevante (patrones, tendencias, correlaciones, etc) para el estudio.

Valores atípicos

Para analizar los valores atípicos, generamos cuatro diagramas de caja y bigotes para las columnas de tipo numérico:



Gráfica 3: Diagramas de cajas y bigotes de las variables numéricas.

En los diagramas de **IngresoAnual**, **Edad** y **CantAutomoviles** observamos valores atípicos. Decidimos que los vamos a tener en cuenta para el análisis en lugar de imputarlos o eliminarlos, esto ya que estos valores nos van a dar información sobre situaciones excepcionales para el estudio. Además, al tenerlos en cuenta podemos ver cómo afectan la forma y la distribución de los datos. También, nuestra conclusión será más precisa al tener en cuenta todos los datos y obtendremos una imagen más precisa de lo que estamos buscando, en este caso potenciales clientes para venderles bicicletas.

Técnicas predictivas

Un modelo predictivo es un modelo analítico que se construye con el objetivo de predecir resultados o comportamientos futuros en función de los datos históricos y los patrones encontrados en ellos.
En este informe, utilizamos tres técnicas predictivas: árbol de decisión, algoritmo de los vecinos más próximos y análisis discriminante lineal. Para cada técnica, la variable a predecir fue **ComproBicicleta** y las variables predictoras fueron solo las numéricas (obviando IdCiudad e IdCliente).
Para poder realizar una comparación entre las tres técnicas y así decidir cual utilizar, generamos una matriz de confusión para cada una. Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

Para todos los casos, entrenamos el modelo con una muestra del 65% de los datos, siendo el 35% restante utilizado para probar el conjunto de prueba (salvo para LDA).

Árbol de decisión

El árbol de decisión es un algoritmo de aprendizaje supervisado que se utiliza para clasificar y predecir valores basados en reglas de decisión en forma de árbol.

La matriz de confusión del árbol de decisión que obtuvimos fue:

accuracy: 70.34%

	true 1	true 0	class precision
pred. 1	596	377	61.25%
pred. 0	287	979	77.33%
class recall	67.50%	72.20%	

Matriz 3: Matriz de confusión del árbol de decisión.

La interpretación de la matriz de confusión es la siguiente:

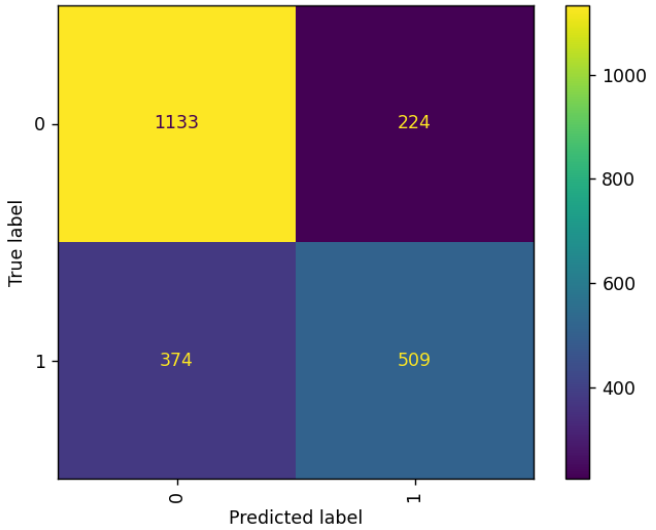
- Se predijo 1 cuando el valor verdadero era 1 (correcto) un total de 596 veces.
- Se predijo 1 cuando el valor verdadero era 0 (incorrecto) un total de 377 veces.
- Se predijo 0 cuando el valor verdadero era 1 (incorrecto) un total de 287 veces.
- Se predijo 0 cuando el valor verdadero era 0 (correcto) un total de 979 veces.

Además, obtuvimos una precisión del 70.34%, lo que implica que alrededor de ese porcentaje de las personas fueron clasificadas correctamente en términos de si compraron o no bicicletas en el pasado. Esta exactitud indica un nivel razonable de rendimiento en la clasificación de las instancias para nuestro conjunto de datos.

Algoritmo de los vecinos más próximos (KNN)

El algoritmo predictivo KNN (K-nearest neighbors) consiste en clasificar un nuevo caso, en función de su distancia con los casos vecinos. Se trabaja con un parámetro “k” que determina la cantidad de vecinos cercanos con los cuales se comparará la nueva observación. En este caso, luego de realizar el respectivo análisis, utilizamos k = 12.

La matriz de confusión obtenida es:



Matriz 4: Matriz de confusión del algoritmo KNN.

La precisión de la matriz de confusión no está en la gráfica, pero se puede calcular fácilmente: es el porcentaje de la suma de las predicciones acertadas sobre el total de valores: *Precisión = (Verdaderos Positivos + Verdaderos Negativos) / Total de casos*.
Luego, la precisión obtenida con el algoritmo KNN es de 73.3%, un poco más que lo obtenido con el árbol de decisión.

Análisis discriminante lineal (LDA)

El Análisis discriminante es una técnica de aprendizaje supervisado utilizada para encontrar una combinación lineal de variables independientes que mejor discrimina entre clases o grupos distintos.

La matriz de confusión obtenida luego del análisis discriminante lineal es:

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		ComproBicicleta	0	1	Total
Original	Recuento	0	2325	1551	3876
		1	921	1603	2524
	%	0	60,0	40,0	100,0
		1	36,5	63,5	100,0

a. 61,4% de casos agrupados originales clasificados correctamente.

Matriz 5: Matriz de confusión del algoritmo LDA.

Observamos una precisión del 61.4%. Esta exactitud es la menor obtenida en relación a las demás técnicas predictivas. Además, la diferencia con respecto a la segunda más baja (70.34%, por parte del árbol de decisión) es mucha. Por lo tanto, descartamos su uso para el estudio.

Conclusión

A pesar de que la precisión del KNN es mayor que el de árbol de decisión (73.3% > 70.34%), decidimos utilizar el árbol de decisión para predecir, debido a:

- La diferencia entre las precisiones no es muy grande (tan solo un 2.96%), por lo tanto, no hay mucha disparidad entre usar un modelo u otro.
- El árbol de decisión predice más falsos positivos que el algoritmo KNN: esto es importante ya que es lo que nos solicitó la jefa de marketing. Es preferible enviarle un correo a una persona que no resulte comprador que perder un potencial cliente.

Predicción

Finalmente, decidimos utilizar el árbol de decisión para la predicción.
Se tiene un total de 1500 destinatarios, de los cuales el árbol predijo:

- No enviarle la publicidad a 823 personas.
- Enviarle la publicidad a 677 personas.