Entrega 2 – Introducción a la I.A

Brayan Daniel Oviedo Barreto, Julián Zaque Montoya, Lisset Andrea Zea
c.c 1037667170, 1152713576, 1001813095
Bioingeniería, Facultad de Ingeniería, Universidad de Antioquia
Abril 23, 2023

DESCRIPCIÓN DEL AVANCE ALCANZADO

EXPLORACIÓN DE LOS DATOS

El dataset a utilizar es el conjunto de datos de la competición "Santander Customer Transaction Prediction" alojado en Kaggle (Santander Customer Transaction Prediction | Kaggle). Este conjunto de datos contiene 200,000 instancias de entrenamiento y 200,000 instancias de prueba, y consta de 200 características numéricas (columnas) anónimas que representan los atributos de los clientes y su historial de transacciones. En estos datos se tiene la variable target la cual es el objetivo a predecir y está compuesta por 1 o 0, lo cual. Teniendo esto en cuenta, se hizo una revisión de los datos:

```
## KEEPOUTPUT
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
d = pd.read_csv("train.csv")
d.head()
```

	ID_code	target	var_0	var_1	var_2	var_3	var_4	var_5	var_6	var_7	...	var_190	var_191	var_192	var_193	var_194	var_195	var_196	var_197	var_198	var_199
0	train_0	0	8.9255	-6.7863	11.9081	5.0930	11.4607	-9.2834	5.1187	18.6266	...	4.4354	3.9642	3.1364	1.6910	18.5227	-2.3978	7.8784	8.5635	12.7803	-1.0914
1	train_1	0	11.5006	-4.1473	13.8588	5.3890	12.3622	7.0433	5.6208	16.5338	...	7.6421	7.7214	2.5837	10.9516	15.4305	2.0339	8.1267	8.7889	18.3560	1.9518
2	train_2	0	8.6093	-2.7457	12.0805	7.8928	10.5825	-9.0837	6.9427	14.6155	...	2.9057	9.7905	1.6704	1.6858	21.6042	3.1417	-6.5213	8.2675	14.7222	0.3965
3	train_3	0	11.0604	-2.1518	8.9522	7.1957	12.5846	-1.8361	5.8428	14.9250	...	4.4666	4.7433	0.7178	1.4214	23.0347	-1.2706	-2.9275	10.2922	17.9697	-8.9996
4	train_4	0	9.8369	-1.4834	12.8746	6.6375	12.2772	2.4486	5.9405	19.2514	...	-1.4905	9.5214	-0.1508	9.1942	13.2876	-1.5121	3.9267	9.5031	17.9974	-8.8104

5 rows × 202 columns

Figura 1. Tabla con las primeras 5 filas y algunas variables que contiene el data set.

Se observaron el nombre de las columnas, el tipo de datos en ellas y la información faltante,

En este paso se observaron diferentes problemas:

1. La cantidad de datos era exagerada teniendo en cuenta los requerimientos mínimos del proyecto y el gasto computacional que estos demandarían.
2. Las variables no tenían nombres que dieran información sobre qué se estaba observando
3. No se contaba con la cantidad mínima de variables categóricas que exigía la entrega.
4. El dataset estaba completo, no contenía variables faltantes. Mientras que los requisitos exigen al menos un 5% de datos faltantes.

Para solucionar estos inconvenientes se llevaron a cabo los siguientes pasos:

- Se recortaron los datos iniciales a un total de 50 columnas y 10000 filas. Este recorte se hizo teniendo en cuenta que los requisitos mínimos eran 30 columnas y 2000 filas, sin embargo, no se quiso limitar estrictamente a esos valores ya que sería una gran pérdida de información.

- Se seleccionaron de manera aleatoria el 10% de las columnas para convertirlas en variables categóricas por medio de la función Kmeans. Se seleccionó como parámetro 5 categorías.
- Se eliminaron los datos de algunas columnas para simular datos faltantes. Esto se hizo por medio de una lista de columnas aleatorias. En este código, 'numpy.random.choice' se usa para seleccionar aleatoriamente tres columnas en el conjunto de datos en la lista 'columnas_faltantes'. Luego, se calcula el número de celdas que deben estar ausentes para cumplir con el requisito del 5% en 'num_celdas_faltantes'. Se seleccionan filas aleatorias en cada columna elegida y se reemplazan los valores correspondientes con NaN. Finalmente, el conjunto de datos modificado se guarda en un nuevo archivo CSV usando la función 'to_csv'.
- Se volvió a analizar el dataset resultante para saber si ahora cumple con las características necesarias.

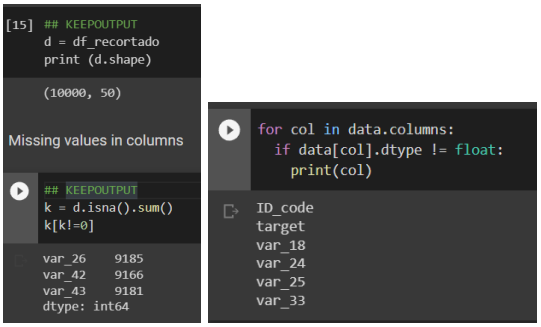

Figura 2. Dimensiones de los datos, datos faltantes en variables y variables categóricas

Al verificar que se cumplían los requisitos básicos, se siguió explorando el data set. Al visualizar la variable objetivo, que en este caso es target, se observó que hay una gran cantidad de 0 (indica que el cliente no ha hecho una transacción) a diferencia de 1 (indica que el cliente ha hecho una transacción)

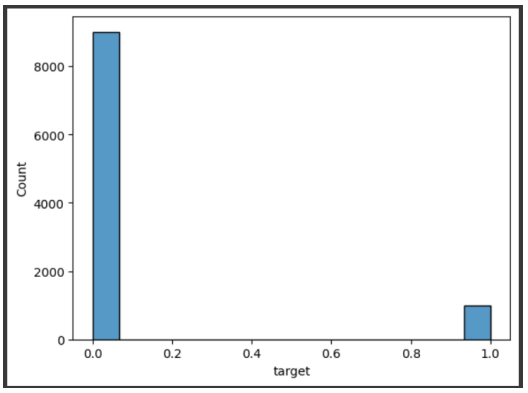

Figura 3. Inspección de la variable objetivo

También se hizo una inspección de algunos como la media, desviación estándar, el máximo, el mínimo, entre otros, de cada una de las variables.

Se observan diferentes desviaciones estándar entre las variables, algunas altas y otras no tanto. Además, una alta diferencia entre las medias. Esto indica una alta variabilidad entre los datos. Una alta desviación estándar en los datos indica que los valores de la muestra están muy dispersos alrededor de la media. En otras palabras, los datos están más alejados de la media y son más heterogéneos. Una baja desviación estándar, por otro lado, indica que los valores están más cercanos a la media y son más homogéneos. Por otro lado, se observa una variabilidad entre las medias de cada variable. La presencia de diferentes medias en un conjunto de datos (data set) indica que hay diferencias o variaciones en los valores que se están midiendo. En otras palabras, los datos no son homogéneos y pueden haber subgrupos o categorías en los datos que están influyendo en las diferentes medias. Se puede decir entonces que la media no es necesariamente representativa de los datos en general.

Por último se hizo una revisión de la correlación entre las variables y la variable objetivo:

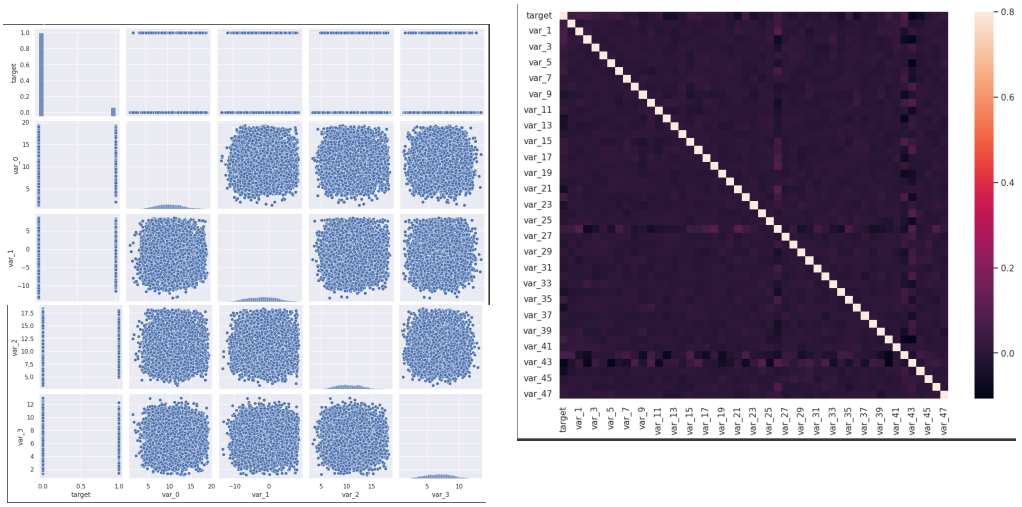

Figura 2. a) Correlación entre variable x y el target. b) Matriz de correlación con todas las variables

En la figura 2 se puede observar que no hay correlación entre la variable target y las demás variables. Esto significa que no hay una relación sistemática o predecible entre ellas. Es decir, los cambios en una variable no están relacionados con los cambios en la otra variable.
Cuando no hay correlación entre dos variables, no hay una relación lineal entre ellas y no se puede utilizar una variable para predecir la otra. Sin embargo, esto no indica que no haya una relación entre las variables. Es posible que haya una relación no lineal o compleja entre las variables que no se puede detectar mediante el análisis de correlación. Por lo tanto, es importante realizar un análisis detallado de los datos y considerar otras técnicas estadísticas para evaluar la relación entre las variables.

No se pudo observar la distribución de las variables categóricas por la cantidad de datos, ya que la imagen más grande que permite es de 2 ^16 pixeles y los datos sobrepasan estas dimensiones.

DATA CLEANING

Posteriormente, se procede a realizar una limpieza de los datos resultantes. En el caso de las variables continuas, se evaluarán 3 técnicas de sustitución para los datos faltantes:

- Reemplazar por el valor de 0
- Reemplazar por el valor del promedio
- Por muestreo de una normal equivalente (misma media y estándar).

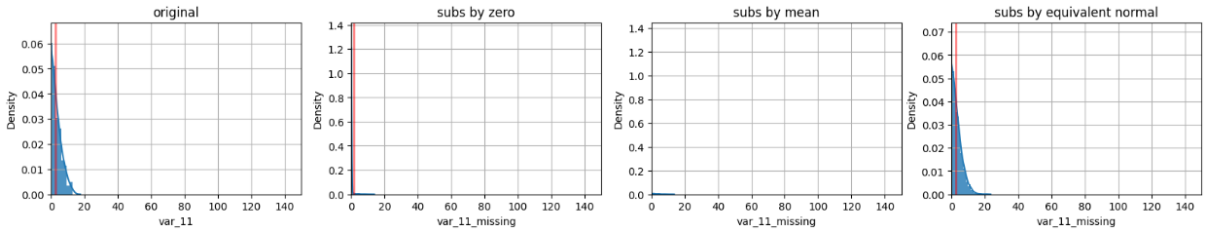

Figura 3. Gráfico de distribución de los datos en la variable 11 al realizar 3 técnicas de sustitución para los datos faltantes.

A priori, al observar la figura 3. se aprecia que con el método de reemplazar los valores faltantes por muestreo de una normal equivalente se obtiene la distribución más similar al comportamiento de los datos originales. Sin embargo, para determinar cuál método de los evaluados anteriormente para reparar datos es mejor se requieren llevar a cabo pruebas de hipótesis , no basta con gráficos de estadística descriptiva. Con base a lo anterior, se crearon modelos predictivos de los datos originales con los datos resultantes de cada técnica de sustitución de datos faltantes y se busca evidencia si los modelos mejoran o no al usar diferentes políticas para reparar datos faltantes. Para esto, se empleó una función llamada HTest que realiza una prueba de hipótesis tipo t-student (ttest_ind) para comparar la precisión de un modelo de regresión aleatoria de bosques (RandomForestRegressor) entre el conjunto de datos de referencia (datos originales) y otros conjuntos de datos (datos obtenidos con cada técnica de reemplazo de datos faltantes). Adicionalmente, también se ajusta el modelo en cada conjunto de datos y se calcular su precisión mediante la validación cruzada y la métrica de error absoluto medio (mean_absolute_error).

```
100% (4 of 4) |########################| Elapsed Time: 0:41:09 Time:  0:41:09
Ttest_indResult(statistic=0.14437172740228532, pvalue=0.8855034241689503)
Ttest_indResult(statistic=0.8145975904482072, pvalue=0.41727781790985286)
Ttest_indResult(statistic=1.0024360686858238, pvalue=0.3186017909119795)
```

El valor de la estadística t se utiliza para evaluar si la diferencia entre las medias de los dos grupos es estadísticamente significativa. Cuanto mayor sea la estadística t, mayor será la diferencia entre las medias de los grupos. El valor p, por otro lado, se utiliza para evaluar la significancia estadística de la estadística t. Si el valor p es menor que un umbral predefinido (normalmente 0.05), se considera que la diferencia entre las medias de los grupos es estadísticamente significativa. En este caso, se están comparando las medias de dos grupos en cada prueba t. El primer par de grupos tiene una estadística t de 0.144 y un valor p de 0.886, lo que indica que no hay una diferencia estadísticamente significativa entre las medias de los grupos. El segundo par de grupos tiene una estadística t de 0.815 y un valor p de 0.417, lo que sugiere nuevamente que no hay una diferencia estadísticamente significativa entre las medias. Finalmente, el tercer par de grupos tiene una estadística t de 1.002 y un valor p de 0.319, lo que sugiere que no hay una diferencia estadísticamente significativa entre las medias de los grupos.

Mediante el uso de la función ‘plot_missing’ se podrán observar la distribución de columnas y la relación con la variable objetivo:

- La función ‘f1’ traza un histograma o gráfico de barras que permite observar la distribución de columnas si es categórica, o en su defecto, un gráfico de densidad si es numérico.
- La función ‘f2’ genera la distribución de la variable objetivo según los valores únicos de la columna categórica, esto, mediante un histograma para observar los valores faltantes.
- La última función ‘f3’ traza un gráfico de densidad para la variable de objetivo si la columna es nula o no.

Mediante la función ‘mlutils.figures_grid’ genera una matriz para observar los gráficos de f1, f2 y f3.

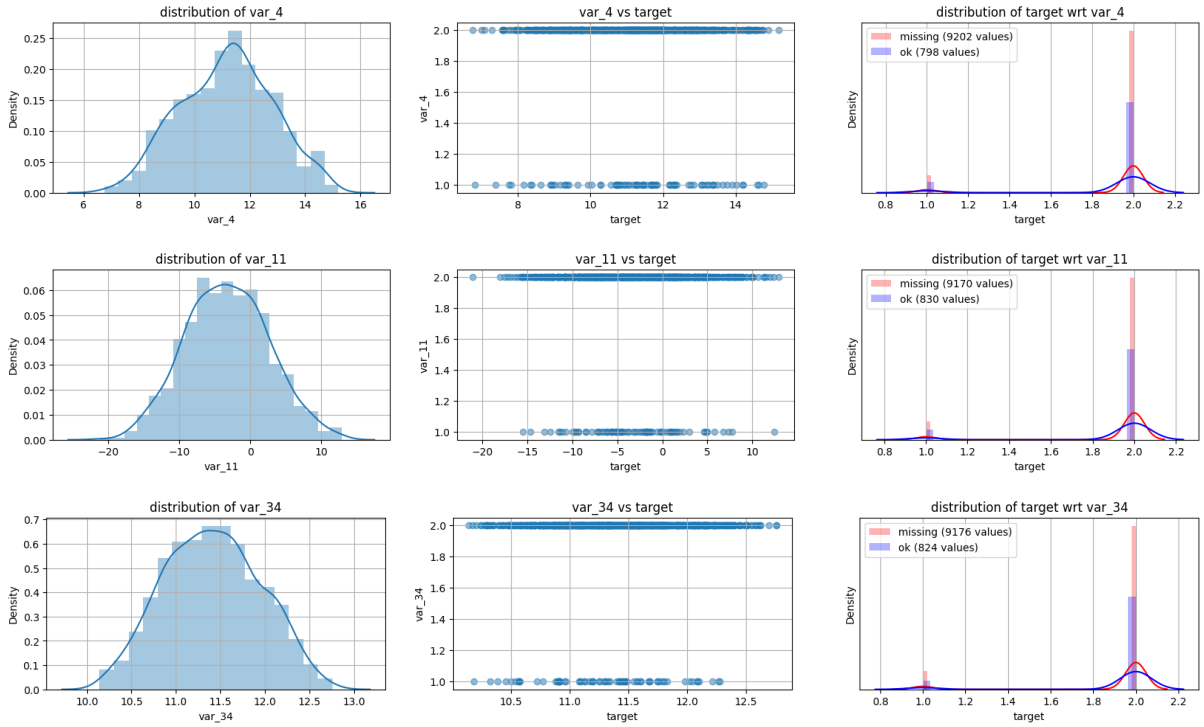

Figura 4. matriz generada por las funciones f1, f2 y f3

CONCLUSIONES

- Se observó que no existe correlación entre las variables y se encontró que no hay una diferencia estadísticamente significativa entre las medias. Esto nos indica que no hay una relación lineal clara entre las variables y la variable objetivo. Esto significa que el modelo no puede predecir la variable objetivo de manera efectiva utilizando solo las variables predictoras. En este caso, es necesario aumentar la cantidad de datos, buscar otras variables predictoras o realizar un análisis más detallado de los datos para encontrar patrones o relaciones más complejas que puedan ayudar a predecir la variable objetivo. Cabe resaltar que es posible que exista una relación no lineal entre las variables predictoras y la variable objetivo. En este caso, técnicas de modelado más avanzadas, como modelos no lineales o de aprendizaje profundo, pueden ser útiles para capturar estas relaciones no lineales y mejorar la precisión del modelo de predicción.
- Al analizar el dataset, se concluyó que tenía una gran cantidad de datos, que en algunos casos se convirtió en un problema para conocer distribuciones y hacer análisis, es por ello que se decidió cortarlo. Sin embargo, no se sabe cuánta de esta información que se eliminó está afectando los resultados de predicción mencionados anteriormente.

Video: https://youtu.be/-ULkgvWmrZc