Optimización de campañas promocionales

a través de la ciencia de datos

CODER HOUSE

Lorenzo Guimaraes

Proyecto Final para Data Science II:
Machine Learning para la ciencia de datos

Análisis e interpretación de caso, visualización y disposición de datos para el desarrollo de un modelo predictivo

CODER HOUSE

[Proyecto en Github]

[Link al Notebook]

Data Science II - Comisión 60895

Lorenzo Guimaraes

Etapas del proyecto

- Identificación de objetivo y desarrollo de hipótesis
- Adquisición de los datos
- Limpieza y ajuste (Data Wrangling)
- Análisis exploratorio
- Tratamiento de los datos:
  Transformación de variables, feature engineering
- Balanceo y escalado
- Principal Component Análisis
- Aplicación de modelos:
  -DecisionTree
  -Regresión logística
  -RandomForestClassifier
- Validación y optimización de modelos
- Comparación y selección de modelo
- Conclusión

La empresa

Una empresa del sector de gastronomía. con más de 100 mil clientes se especializa en la venta de alimentos gourmet, carnes, vinos, frutas exóticas y demás productos, a través de tres medios diferentes, via web, en la tienda física, o catálogos.

En este caso analizaremos una campaña piloto realizada con el fin de realizar un muestreo del alcance de su campañas promocionales a fin de maximizar el impacto de campañas futuras junto al lanzamiento de sus nuevos productos.

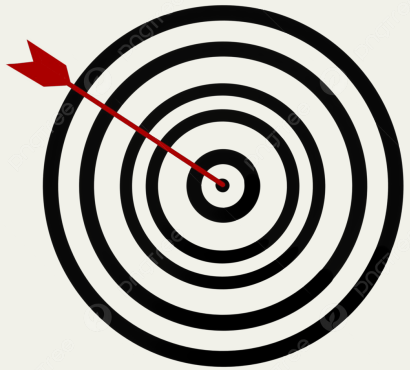
La campaña

La campaña tuvo 2240 clientes seleccionados al azar a los cuales se ofreció una oferta promocional de sus nuevos productos, los que fueron aceptando en los meses siguientes fueron categorizados como "Si" a la respuesta.

El costo total de la campaña superó casi por el doble a los ingresos generados con las ventas de la oferta promocional.

La empresa cree que los pronósticos para los próximos tres años no son muy prometedores... por lo que quieren hacer algo al respecto.

Descripción y objetivo

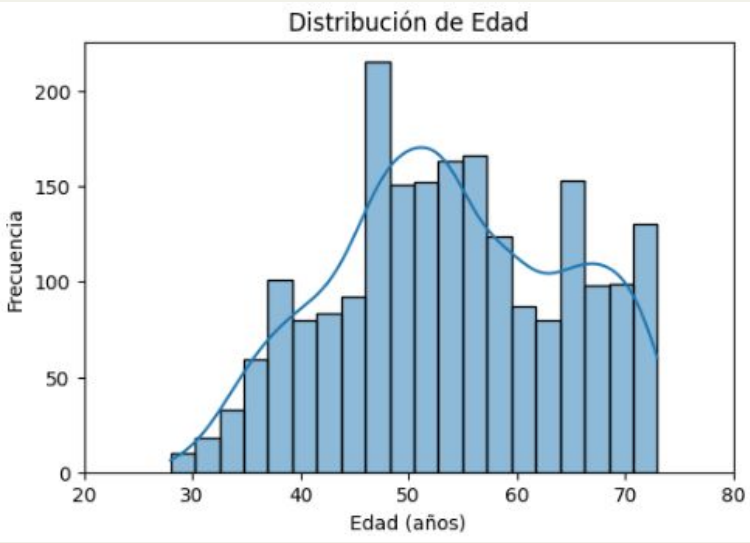


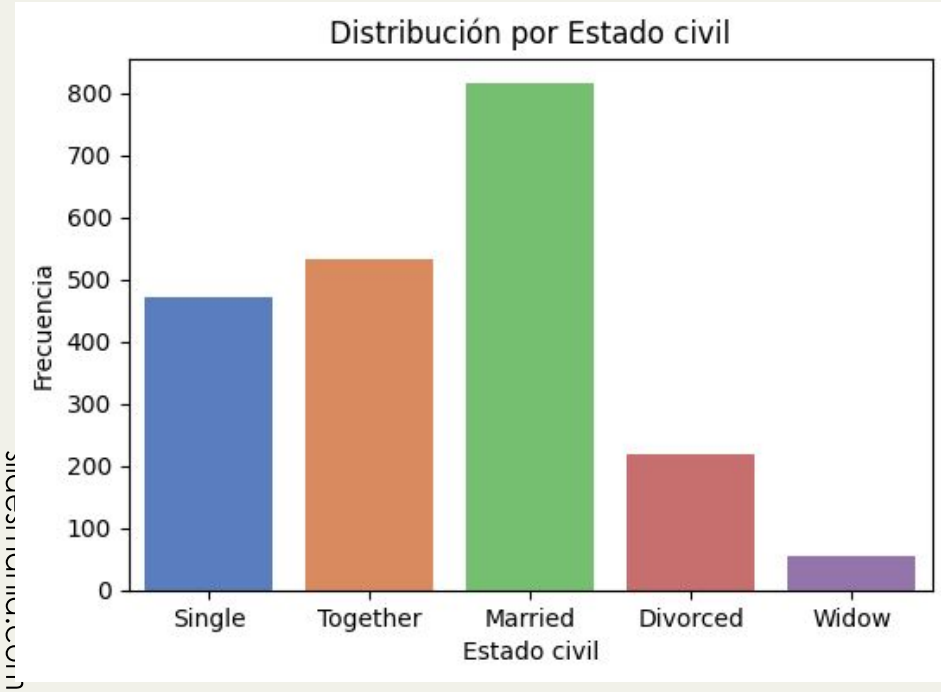
El objetivo de este proyecto es desarrollar un modelo predictivo el cual permita predecir el comportamiento de los clientes y aplicarlo al resto de la customer base.

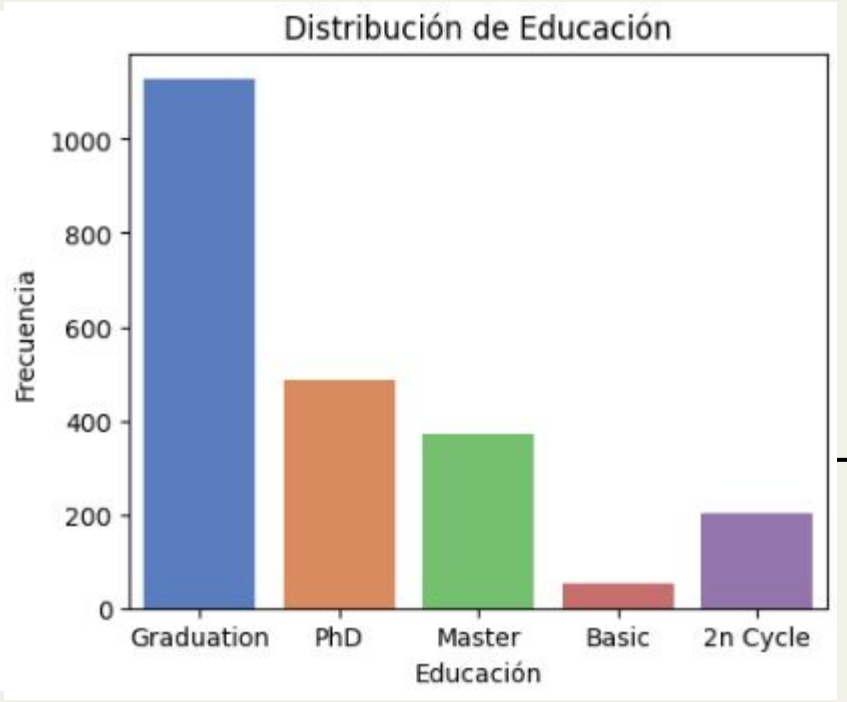
Se busca maximizar el impacto de las futuras campañas y lograr ofrecer promociones seleccionando específicamente a los clientes que aceptarían la oferta, y evitar a los que la rechazarían, así hacer la campaña altamente rentable.

Algunas visualizaciones sobre los clientes

Algo que resultó no tener mucha correlación con la respuesta a la campaña promocional.

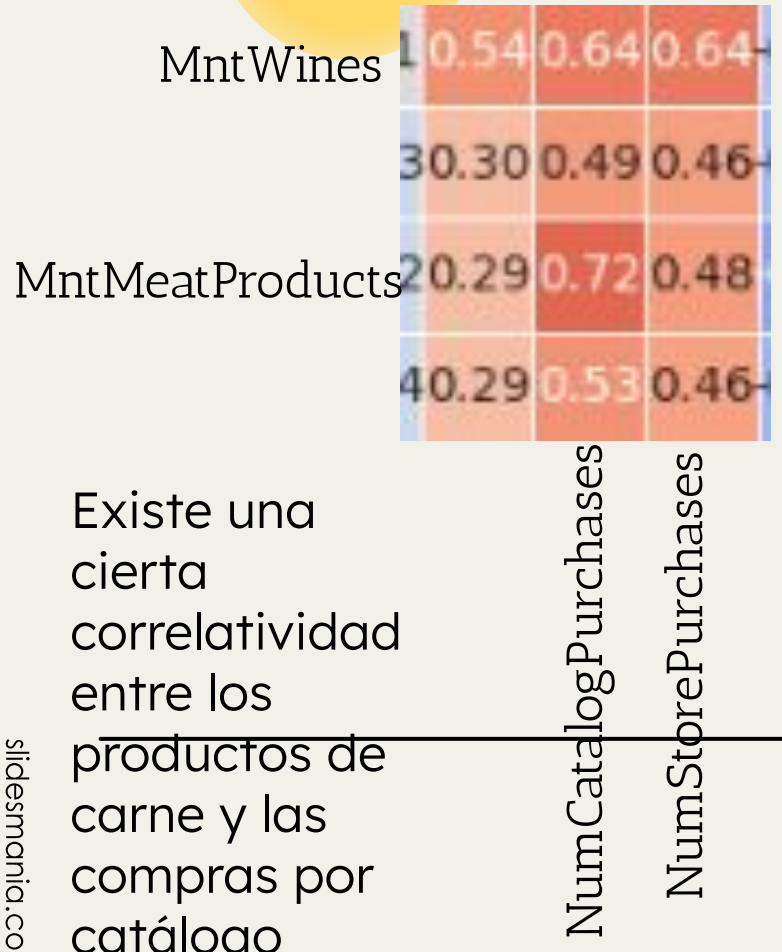




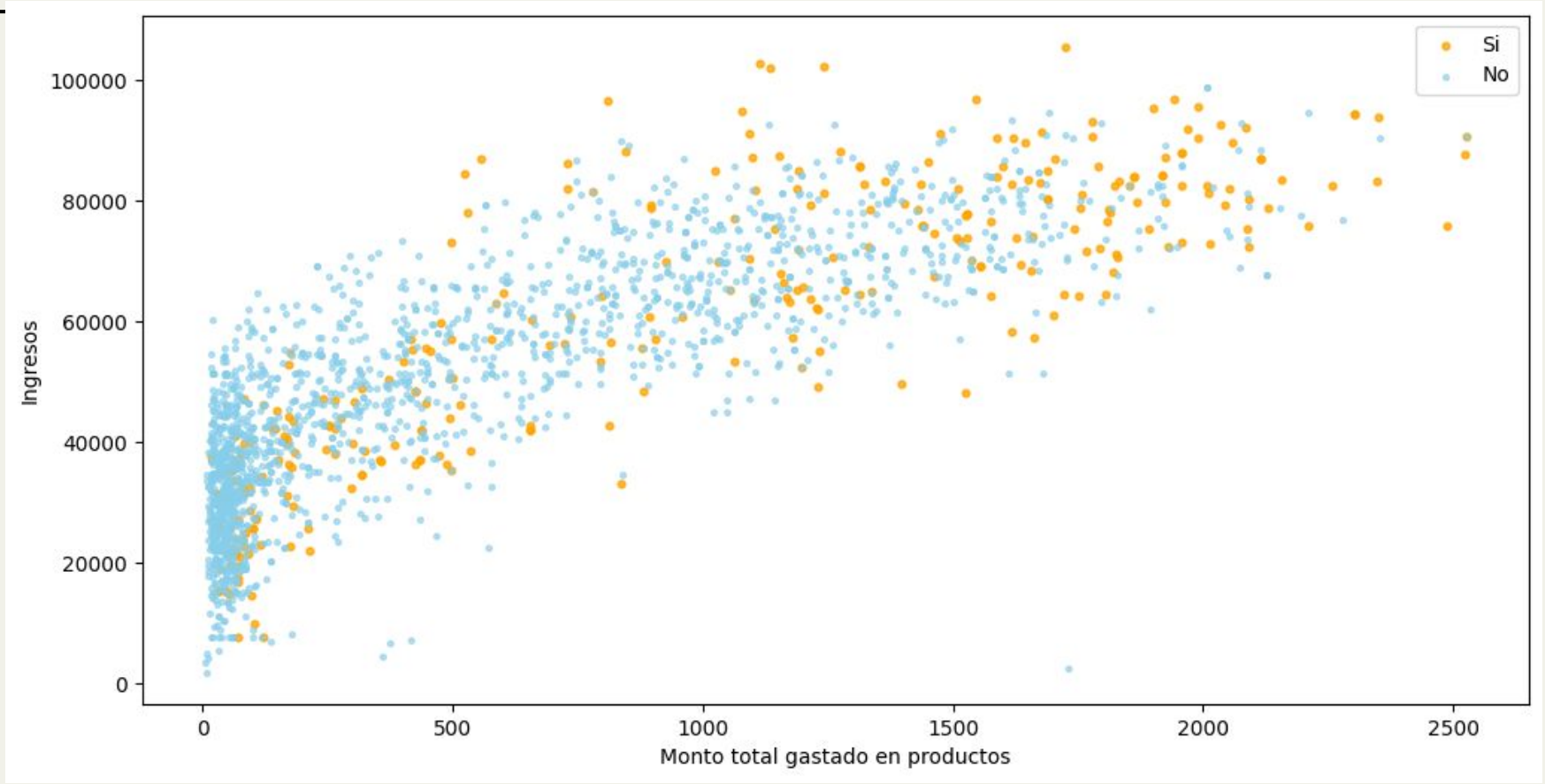


Aunque si debe haber una posibilidad de segmentación de clientes para armar grupos target.

Correlaciones



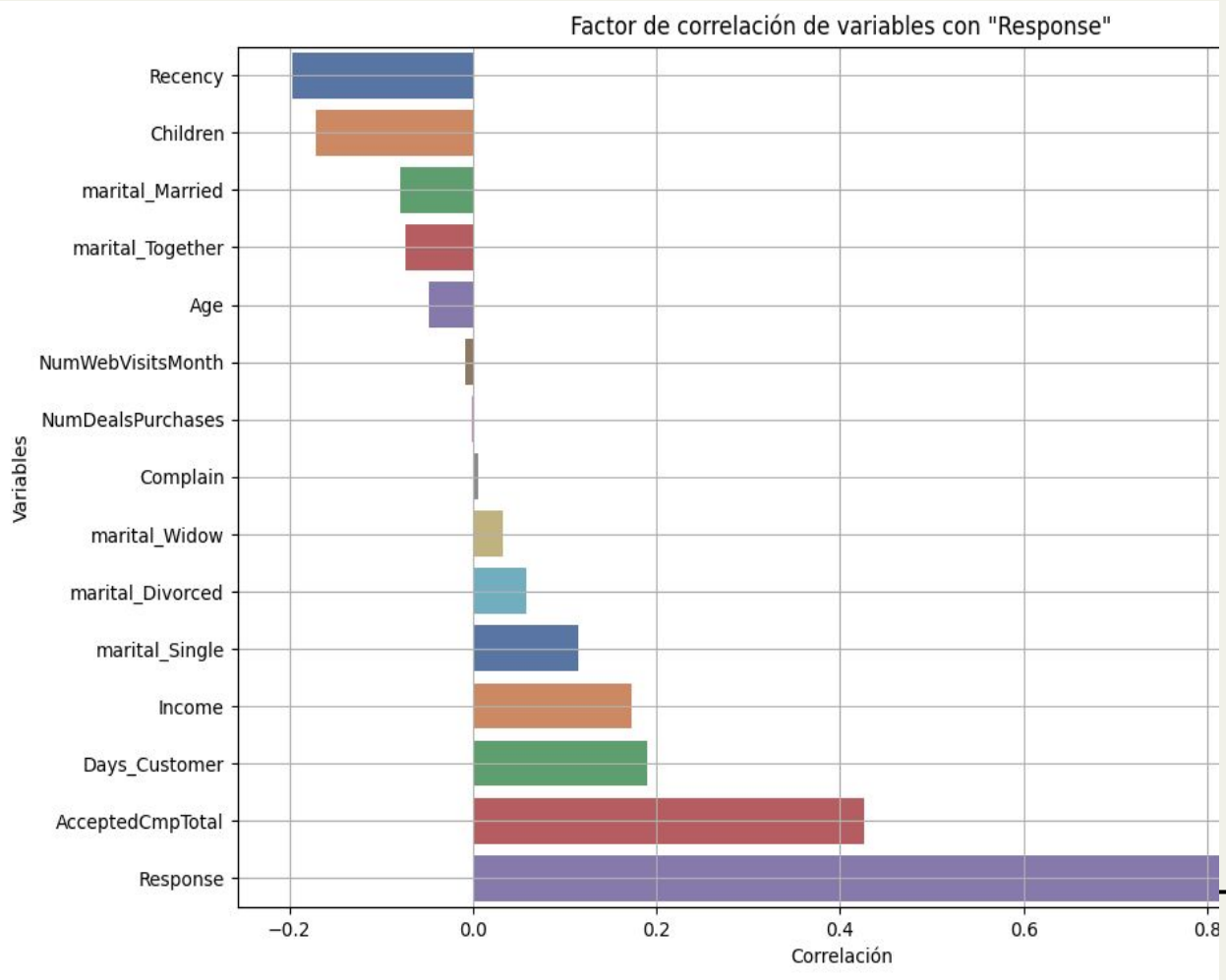
Existe una cierta correlatividad entre los productos de carne y las compras por catálogo

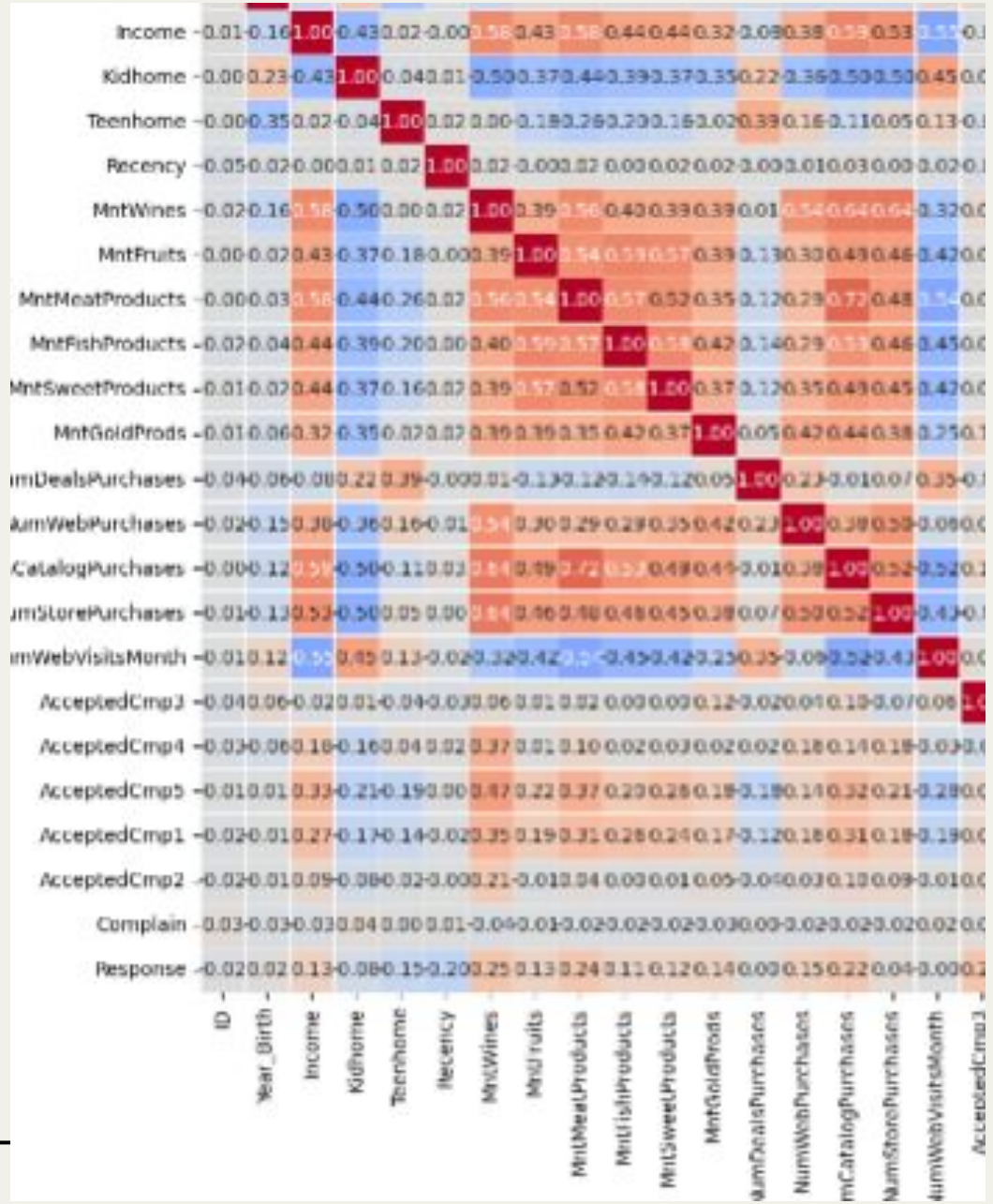


La correlación entre Ingresos y monto total gastado en productos claramente tiene una pendiente positiva.

Visualización de correlaciones

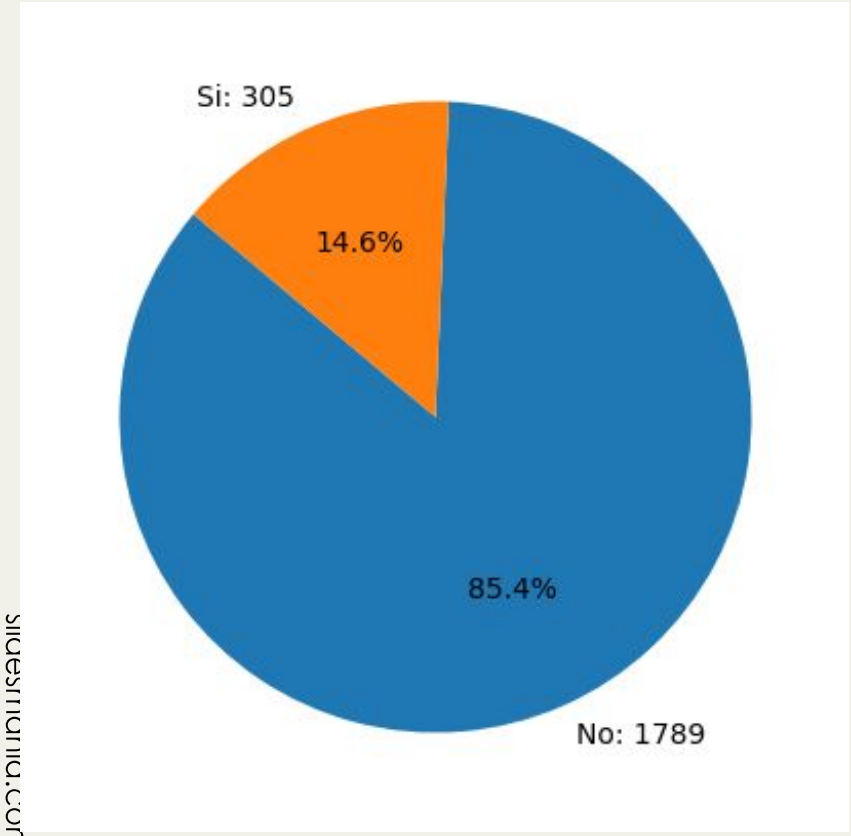
El factor de correlación (entre -1 y 1) visto en barras o el heatmap dan una buena perspectiva de correlación entre columnas, y tambien son indicativas de dependencia entre unas y otras





Balanceo de muestras

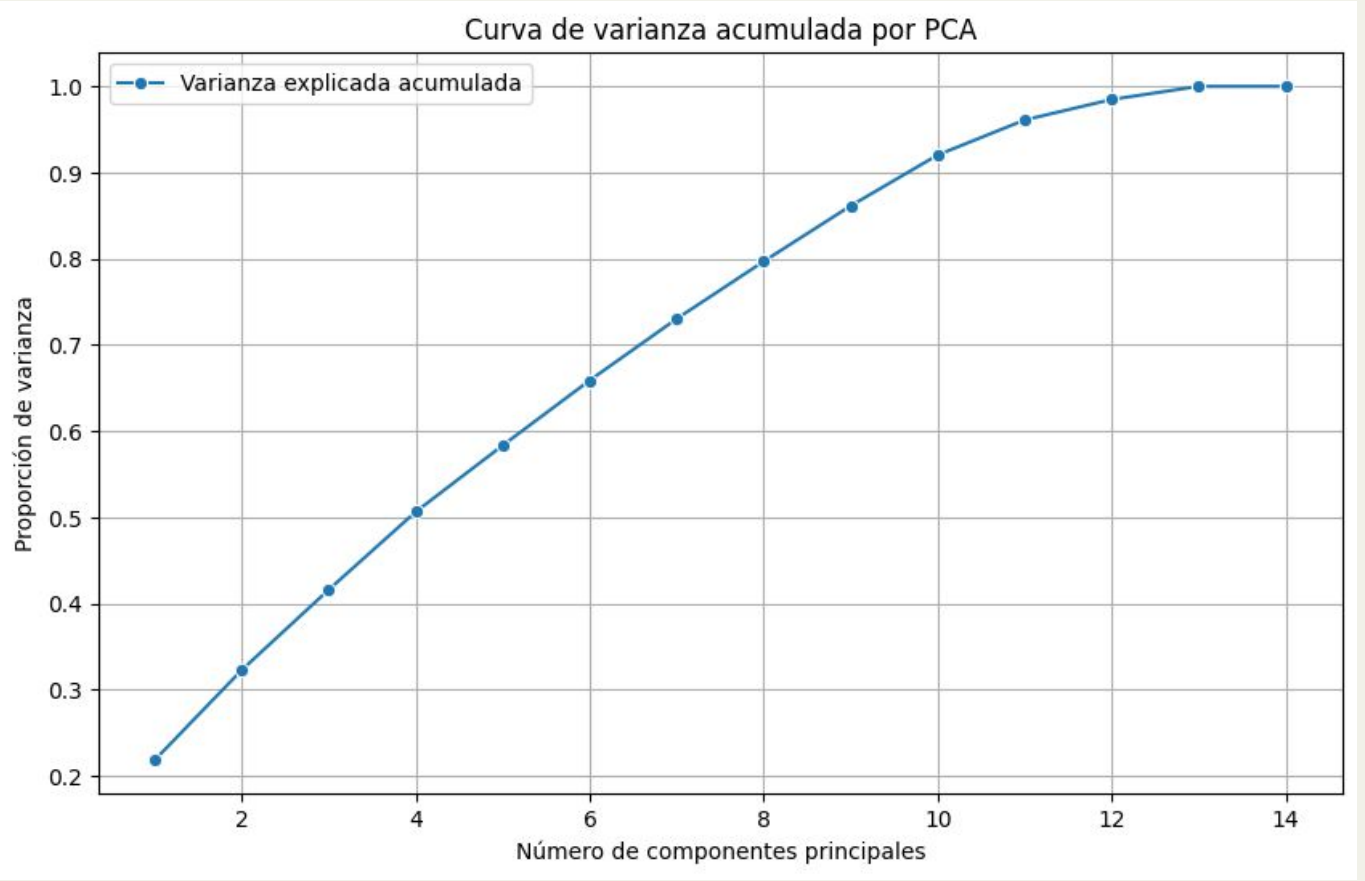
Al tener una distribución desigual en las muestras de la variable a predecir, puede traer un desbalance en el entrenamiento, lo cual genera dificultades para la predicción.



Se hizo un balanceo con la técnica de oversampling, lo cual produce muestras sintéticas en base a los datos existentes, para igualar la cantidad.

PCA – Análisis de componentes principales

El PCA es una tecnica la cual mide la varianza que aporta la cantidad de componentes.
El punto ideal en mi caso está en los 12 componentes, donde ya no hay un aumento significativo.
Cuanto mayor la acumulacion de varianza y menor la cantidad de componentes, mejor.



Modelos de clasificación

Se eligieron estos 3 modelos para entrenar con datos balanceados y escalados, a los cuales se les aplicó el PCA, el cuál se dejó a un lado ya que no tuvo buenos resultados.

Se realizó una validación simple con una matriz de confusión, y Accuracy, Precision, Recall, y F1.
Se hizo una validación cruzada con StratifiedKFold, la cual se ve en la siguiente diapositiva.

>Regresión Logística

Accuracy: 0.7947
Precision: 0.4051
Recall Score: 0.7343
f1 Score: 0.5222

>Arbol de decisión

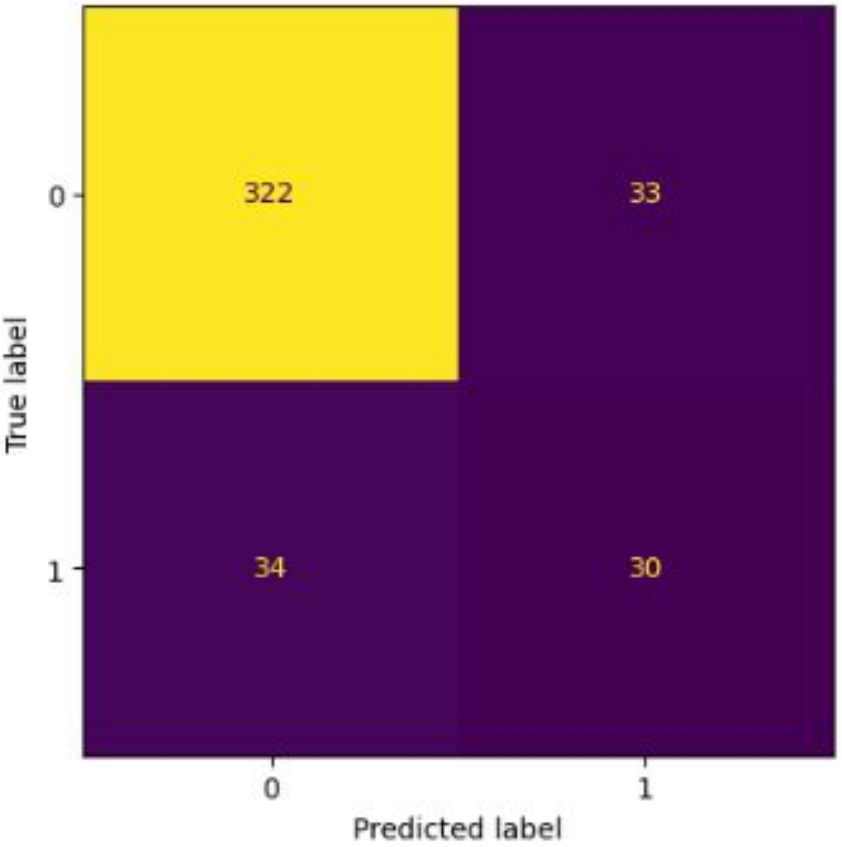
Accuracy: 0.8400
Precision: 0.4761
Recall: 0.4687
F1 Score: 0.4724

>RandomForestClassifier

Accuracy: 0.8878
Precision: 0.6888
Recall Score: 0.4843
f1 Score: 0.5688

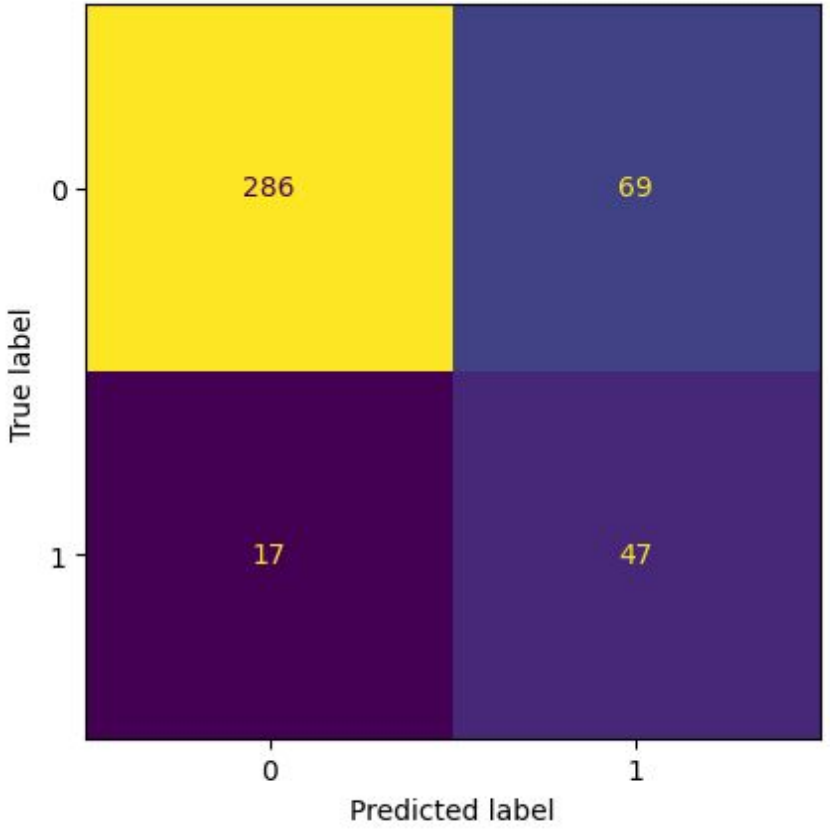
Validación simple y cruzada

Decision Tree



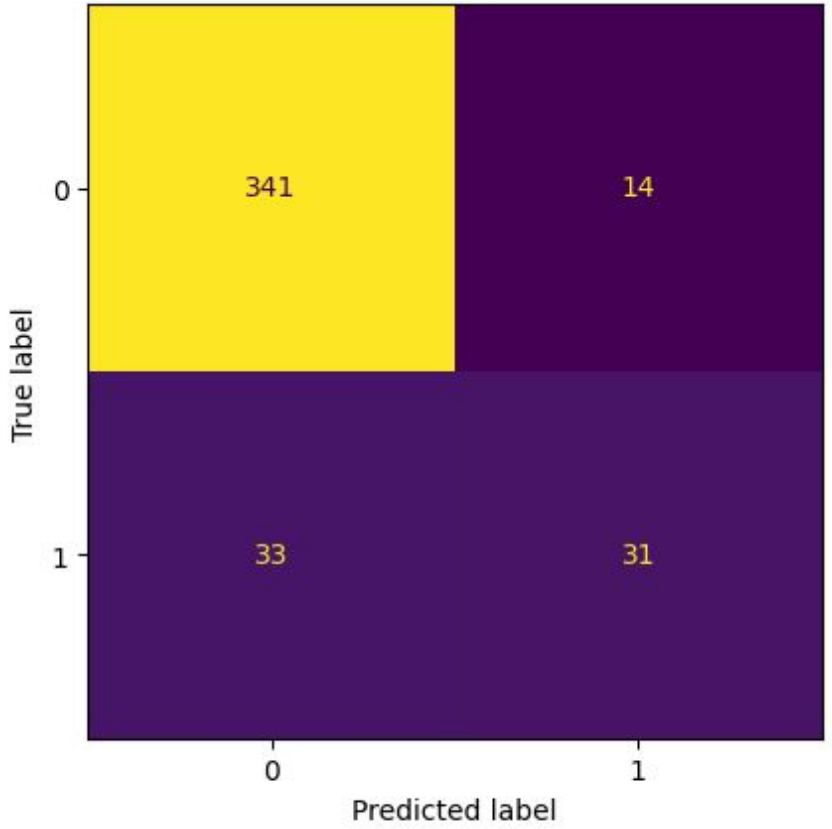
StratifiedKFold Results:
Accuracy: 0.7846
Precision: 0.3687
Recall: 0.6623
F1 Score: 0.4733

Regresión Logística



StratifiedKFold Results:
Accuracy: 0.7798
Precision: 0.3754
Recall: 0.7541
F1 Score: 0.5008

RandomForestClassifier



StratifiedKFold Results:
Accuracy: 0.8945
Precision: 0.7677
Recall: 0.3934
F1 Score: 0.5191

Optimización de los modelos

Para optimizar los modelos, se utilizo el GridSearchCV para ajustar los hiperparámetros de los algoritmos de clasificación.

Decision Tree

Falsos negativos: 8.11% (34/419)
Falsos positivos: 7.88% (33/419)

Regresión Logística

Falsos negativos: 4.06% (17/419)
Falsos positivos: 16.47% (69/419)

RandomForestClassifier

Falsos negativos: 7.88% (33/419)
Falsos positivos: 3.34% (14/419)

Resultados luego de optimizar con GridSearchCV

Accuracy: 0.8067
Presición: 0.4045
Recal: 0.5625
F1: 0.4706

Accuracy: 0.8043
Presició: 0.4167
Recal: 0.7344
F1: 0.5341

Accuracy: 0.8878
Presición: 0.6735
Recall: 0.5156
F1: 0.5841

Comparación de modelos y conclusiones

	Modelo	Accuracy	Precisión	Recall	F1 Score
0	Regresión Logística	0.804296	0.419643	0.734375	0.534091
1	Árbol de Decisión	0.806683	0.404494	0.562500	0.470588
2	Random Forest	0.887828	0.673469	0.515625	0.584071

El modelo que mejores resultados tuvo fue claramente el RandomForestClassifier, ya que logro tener un recall relativamente alto, manteniendo una muy buena accuracy y precision; métricas clave para nuestro caso, en el cual la precision es escencial para encontrar los casos positivos, que si comprarian de la oferta; y descartar desde un comienzo a los casos negativos, para no perder recursos.

Gracias!

Proyecto en Github

Lorenzo Guimaraes
Data Science II - Comisión 60895

CODER HOUSE

Link al Notebook

Fuente y Autoría

- Fuente de Datos: Marketing Campaign - Kaggle
- Referencia: O. Parr-Rud. *Business Analytics Using SAS Enterprise Guide and SAS Enterprise Miner*. SAS Institute, 2014.