PROYECTO FINAL
DATA SCIENCE

Autor:
Jean Piero Marcon

Institución:
Coderhouse

Fecha de presentación:
10/2022

CODER HOUSE

Contenido

1. Descripción del negocio y de la finalidad del estudio

El análisis de los clientes ayuda a modificar un producto en función a los diferentes segmentos de ellos. Por ejemplo, desarrollar estrategias de ventas enfocadas a los clientes de mayor potencial, y con productos específicos.

En la siguiente sección, iremos a través de un proyecto de ciencia de datos sobre el análisis de los clientes con python, buscando predecir aquellos que sean potencialmente aceptadores de las campañas de marketing, todo esto con datos recopilados de datos personales y desde las ventas para tener un amplio panorama del comportamiento.

2. Diccionario de los datos

De la persona

1. ID: Identificador de cliente
2. Year_Birth: Año de nacimiento
3. Education: Nivel de educación
4. Marital_Status: Estado civil
5. Income: Ingreso familiar anual
6. Kidhome: Número de niños en el hogar
7. Teenhome: Número de adolescentes en el hogar
8. Dt_Customer: Fecha de alta en la empresa
9. Recency: Número de días desde la última compra
10. Complain: 1 si el cliente tuvo quejas en los últimos 2 años, 0 si no

Del producto

11. MntWines: Gasto en vino en los últimos 2 años
12. MntFruits: Gasto en frutas en los últimos 2 años
13. MntMeatProducts: Gasto en carne en los últimos 2 años
14. MntFishProducts: Gasto en pescado en los últimos 2 años
15. MntSweetProducts: Gasto en dulces en los últimos 2 años
16. MntGoldProds: Gasto en oro en los últimos 2 años

De las promociones

17. NumDealsPurchases: Numero de compras hechas con descuento
18. AcceptedCmp1: 1 si el cliente acepto la oferta en la 1ra promoción, 0 si no
19. AcceptedCmp2: 1 si el cliente acepto la oferta en la 2da promoción, 0 si no
20. AcceptedCmp3: 1 si el cliente acepto la oferta en la 3ra promoción, 0 si no
21. AcceptedCmp4: 1 si el cliente acepto la oferta en la 4ta promoción, 0 si no
22. AcceptedCmp5: 1 si el cliente acepto la oferta en la 5ta promoción, 0 si no
23. Response: 1 si el cliente acepto la oferta de la última promoción, 0 si no

Del modo de compra

24. NumWebPurchases: Número de compras hechas a través de la página web
25. NumCatalogPurchases: Número de compras hechas a través de catalogo
26. NumStorePurchases: Número de compras hechas en tienda
27. NumWebVisitsMonth: Número de visitas web en el último mes

3. Herramientas tecnológicas utilizadas

- Anaconda y Google Colaboratory: Para la creación y manipulación de archivos Jupyter Notebook con Python. A la vez, utilizando librerías dentro de python, como Numpy, Pandas, Matplotlib, Seaborn y Sklearn para el entrenamiento y evaluación de los modelos de Machine Learning.
- Ilustrator y Canva: Para la creación y edición de imágenes.
- Microsoft Word: Para la creación de informes.
- Kaggle: Base de datos de datasets funcional para ejemplos de investigación.

4. Dataset a utilizar

El conjunto de datos a estudiar fue sacado de la página web https://raw.githubusercontent.com/amankharwal/Website-data/master/marketing_campaign.csv, como un archivo sin formato .csv para luego ser codificado en python y trabajado como un dataframe.

5. Exploración de los datos

Se consulto el set de datos obteniendo información como sus dimensiones de 2240 registros y 29 datos por registro, la cantidad datos vacíos conteniendo 24 entradas nulas que representa solo un 1% de todo el conjunto, y un resumen general donde se observó para una visualización general e introductoria información estadística como el promedio, la desviación estándar, los cuartiles y valores mínimos y máximos por cada columna.

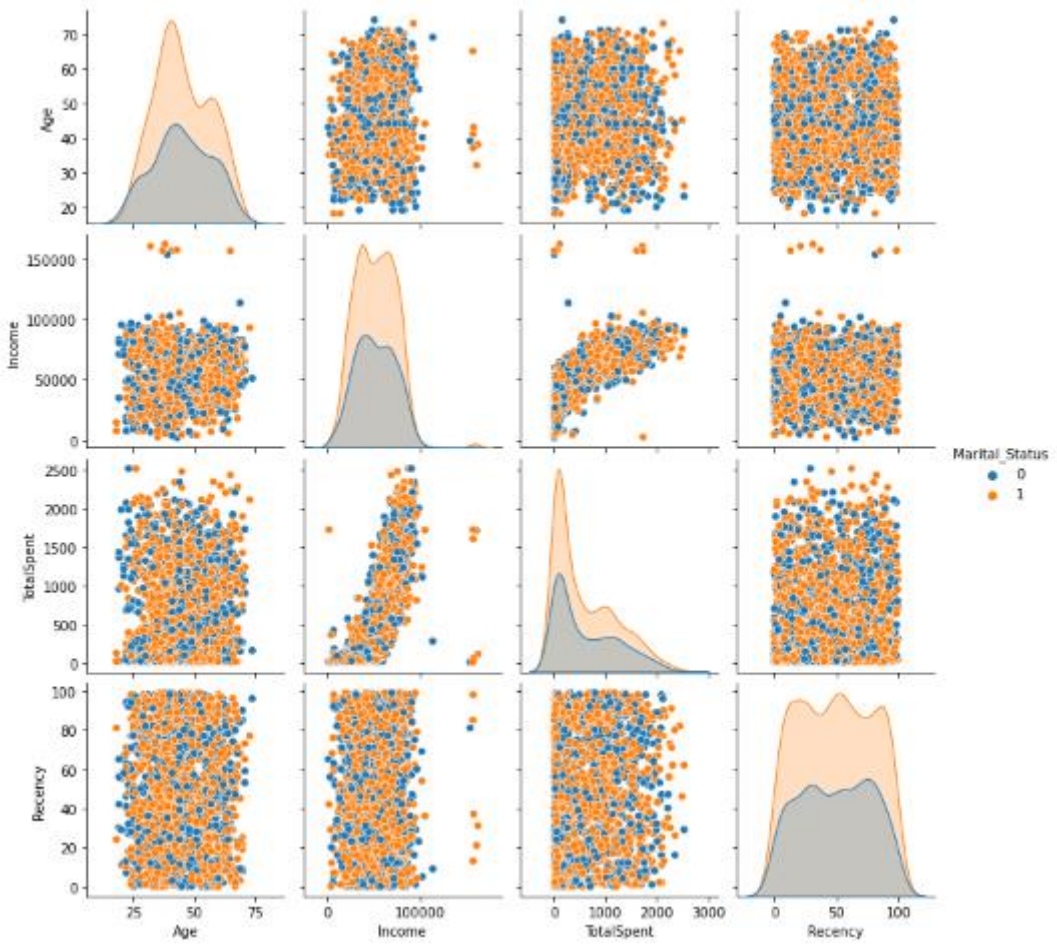
6. Procesamiento de los datos

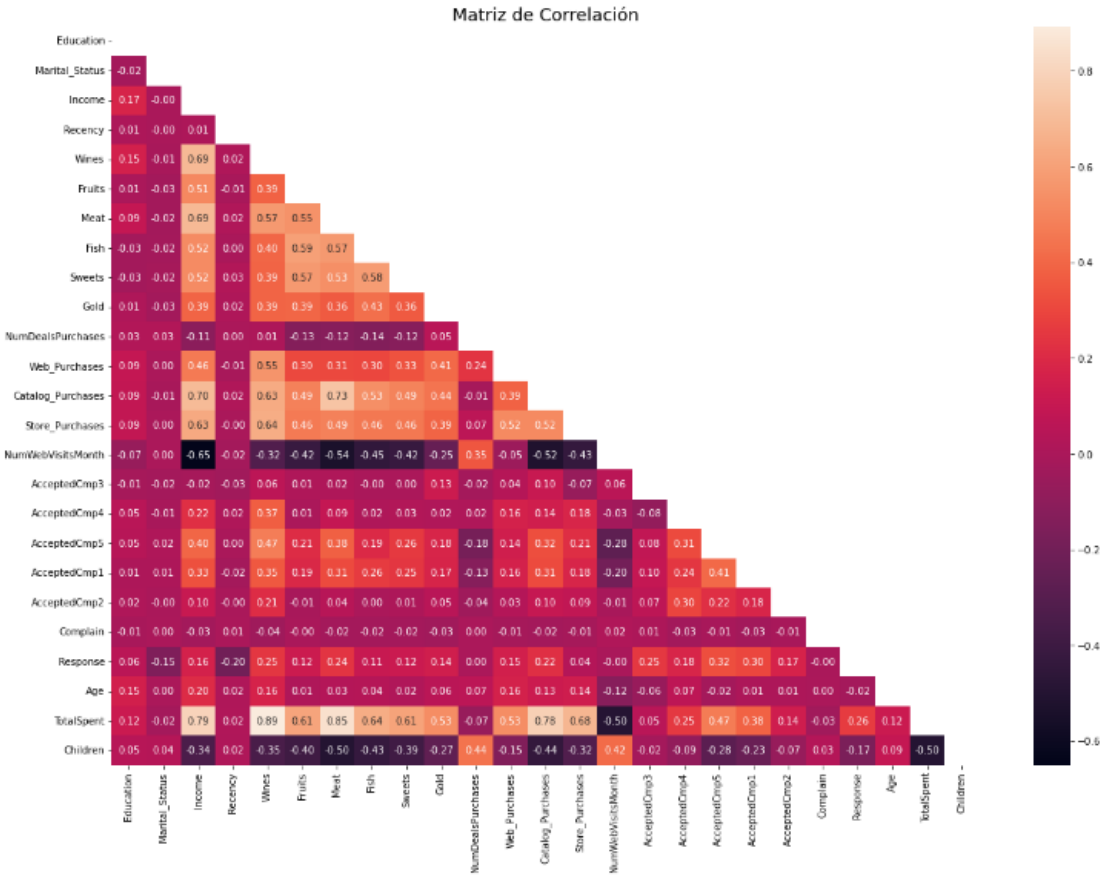
Se hizo una adecuación de los datos para su mejor manipulación y desempeño óptimo de los algoritmos a entrenar, entre los cuales está la eliminación de los registros que contenían datos vacíos ya que representaba una muy baja cantidad dentro de toda la muestra, se eliminaron datos considerados anormales que no representan peso en la muestra y por el contrario podían alterar los resultados erróneamente, se eliminaron columnas sin grandes significados, se calcularon métricas a partir de otros datos dejándolos más útiles y entendibles, y se llevaron los datos categóricos a numéricos normalizándolos.

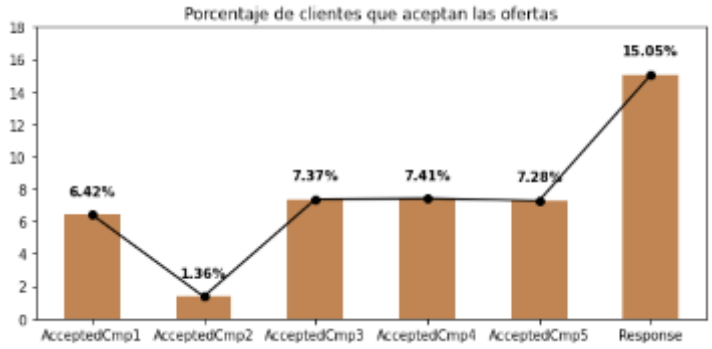
7. Análisis exploratorio de los datos

Se realizaron análisis de correlación de las variables en búsqueda de insights llegando a las siguientes conclusiones:

- Clientes con alto ingreso son los que mas gastan
- Los productos “Meat” y “Wine” son los de mas alto costo
- Clientes con niños y bajo ingreso prefieren las compras en ofertas
- La ultima campaña de marketing tuvo la mayor aceptación respecto a las anteriores







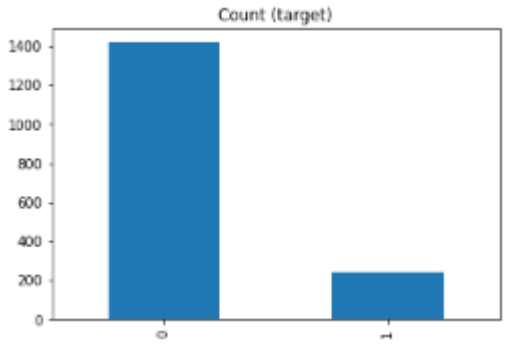
8. Entrenamiento de modelos supervisados

En principio se realizó una separación de los datos escogidos para entrenamiento y los datos del target, y a su vez los datos de entrenamiento en un 75% y los de testeo en el 25% restante, estableciendo un punto de aleatoriedad fijo, obteniendo 4 conjuntos de datos.

```
#Seleccionar datos a entrenar y datos de salida
x = data[['Education', 'Marital_Status', 'Income', 'Children', 'Age', 'TotalSpent', 'Recency']]
y = data['Response']

# División de datos de entrenamiento y datos de prueba
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=42)
```

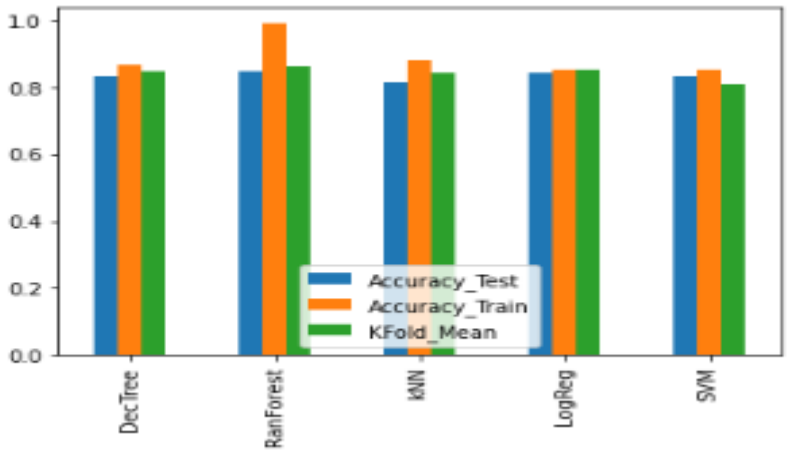
Luego se encontró un desbalanceo en los datos de entrenamiento con un 85,5% en la clase mayoritaria de los negativos, y solo un 14,5% en la clase minoritaria de los positivos siendo la principal del estudio, sugiriendo la aplicación de un método de resampling para equilibrar la muestra.



	counts	%
0	1418	0.854732
1	241	0.145268

Primero se hizo el entrenamiento de 5 modelos en sus versiones predeterminadas para hacer sus comparaciones y evaluaciones con diferentes métricas de validación, obteniendo como mejor opción el algoritmo de Regresión Logística por su alta eficacia y baja diferencia entre la precisión del set de train y el set de test.

	DecTree	RanForest	kNN	LogReg	SVM
Accuracy_Test	0.833635	0.849910	0.817360	0.842676	0.833635
Accuracy_Train	0.870404	0.992164	0.880048	0.852923	0.854732
KFold_Mean	0.848082	0.864351	0.844012	0.855322	0.811964
KFold_Std	0.032013	0.030547	0.029678	0.022423	0.027430



9. Optimización de parámetros

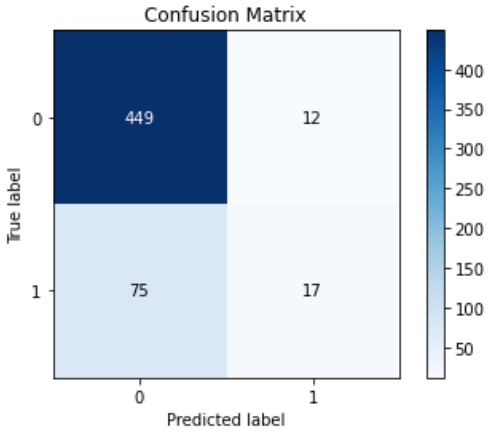
Luego se hizo una optimización de parámetros de los algoritmos buscando las mejorías posibles en las métricas de cada uno, utilizando la los métodos de búsqueda de parámetros por grilla y la búsqueda de parámetros aleatorios.

Con los resultados obtenidos se consiguió una estandarización del accuracy en todos los modelos acercándolos al 83%, lo que fue en mejoría de algunos como en decaimiento en otros, pero siguiendo con mejores métricas el algoritmo de Regresión Logística en su versión predeterminada como lo habíamos optado anteriormente.

	DecTree	RanForest	kNN	LogReg	SVM
Accuracy_Test	0.833635	0.84991	0.81736	0.842676	0.833635
Accuracy_Train	0.870404	0.992164	0.880048	0.852923	0.854732
KFold_Mean	0.848082	0.864351	0.844012	0.855322	0.811964
KFold_Std	0.032013	0.030547	0.029678	0.022423	0.02743
GridSearch	0.835443	0.83906	0.83906	0.83906	0.833635
RandomSearch	0.835443	0.842676	0.835443	0.703436	0.833635

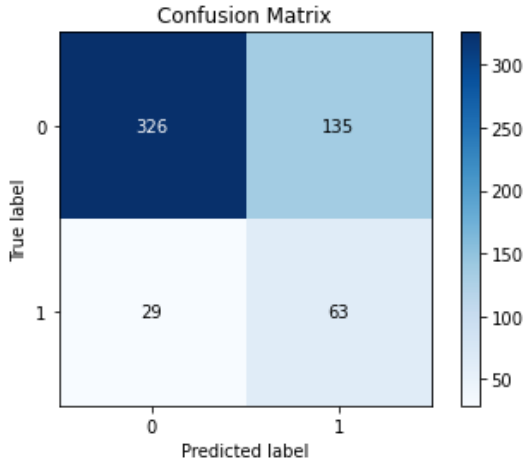
10. Evaluación del modelo seleccionado

Se realizó una matriz de confusión para comparar la asertividad del algoritmo seleccionado en base a la clasificación de la variable a predecir, junto a sus métricas, donde el recall nos muestra nuevamente el desbalanceo de los datos basado en una baja predicción de la clase positiva arrojando un 82% de error.



	precision	recall	f1-score	support
0	0.86	0.97	0.91	461
1	0.59	0.18	0.28	92
accuracy			0.84	553
macro avg	0.72	0.58	0.60	553
weighted avg	0.81	0.84	0.81	553

Esto nos obliga a utilizar una de las técnicas utilizadas para trabajar con el desbalanceo de clases llamada penalización, donde se agregó un parámetro en el modelo de ensamble Regresión Logística, y con esto el mismo algoritmo se encarga en el entrenamiento de equilibrar las clases bajando la cantidad de datos de la clase con mayor cantidad.



	precision	recall	f1-score	support
0	0.92	0.71	0.80	461
1	0.32	0.68	0.43	92
accuracy			0.70	553
macro avg	0.62	0.70	0.62	553
weighted avg	0.82	0.70	0.74	553

Observamos en los resultados un incremento en la detección de verdaderos positivos, pero sacrificando el porcentaje de asertividad de verdaderos negativos, sin embargo para los fines del estudio donde se busca la predicción de los positivos se toma como mejor desempeño.

11. Conclusión

Se construyó un algoritmo de Regresión Logística para predecir los clientes potenciales en aceptar las campañas de marketing ofrecidas por la empresa, con una asertividad del 70% la cual se puede mejorar usando otras técnicas de balanceo del set de datos.

Luego se recomienda usar un algoritmo no supervisado así segmentar los clientes por sus características y lograr enfocar de mejor manera y sin malgastar esfuerzos las estrategias de ventas conociendo ya los perfiles a los cuales acceder, y el uso de ambos algoritmos en conjunto subirá en gran medida esta eficacia.