Data Scientist **Hackathon**

Realizando un modelo predictivo de dos años de pérdida de clientes (churn)
Producto 'Kin Safety' - 2019

RICARDO BAUTISTA HUERTA

Obteniendo población deseada

Para obtener la población objetivo se realizaron 5 filtros :

- Solo se consideraron los contratos del 2015 en adelante: **623242 registros**
- Se descartaron las operaciones en Italia este año (2019): **560947 registros**
- Solo se consideraron los clientes con sólo un contrato: **544615 registros**
- Se eliminaron los clientes que contarán con más del 75% de su información faltante: **526496 registros**
- Solo se consideraron clientes que tuviesen, al menos, 2 años de información en la empresa: **491796 registros**

Tras aplicar los filtros, la población deseada asciende a 491796 registros

	CustomerId	Surname	Geography	Gender	HasCrCard	IsActiveMember	EstimatedSalary	application_date	exit_date	birth_date
0	15745584	EIRLS	Germany	Female	0.0	1.0	0.00	2018-12-14	NaT	1997-09-18
3	14648573	NALLS	Spain	Male	1.0	0.0	140827.98	2019-06-19	NaT	1979-02-27
5	15638124	BRASHERS	Italy	Female	0.0	0.0	170661.45	2018-02-23	NaT	1983-01-13
14	15165393	LABIANCA	Spain	Male	1.0	1.0	2612.65	2018-02-22	2019-06-11	1974-07-11
15	14611239	DOKKA	France	Male	0.0	1.0	72210.60	2019-02-24	NaT	1986-04-26
...	...	...	...	...	...	...	...	...	...	...
1544982	15923060	EISENSTEIN	Germany	Female	1.0	1.0	3627.11	2018-09-24	NaT	1977-01-21

Otras variables relevantes

De las otras tablas se obtuvieron las demás variables para aplicar el modelo predictivo:

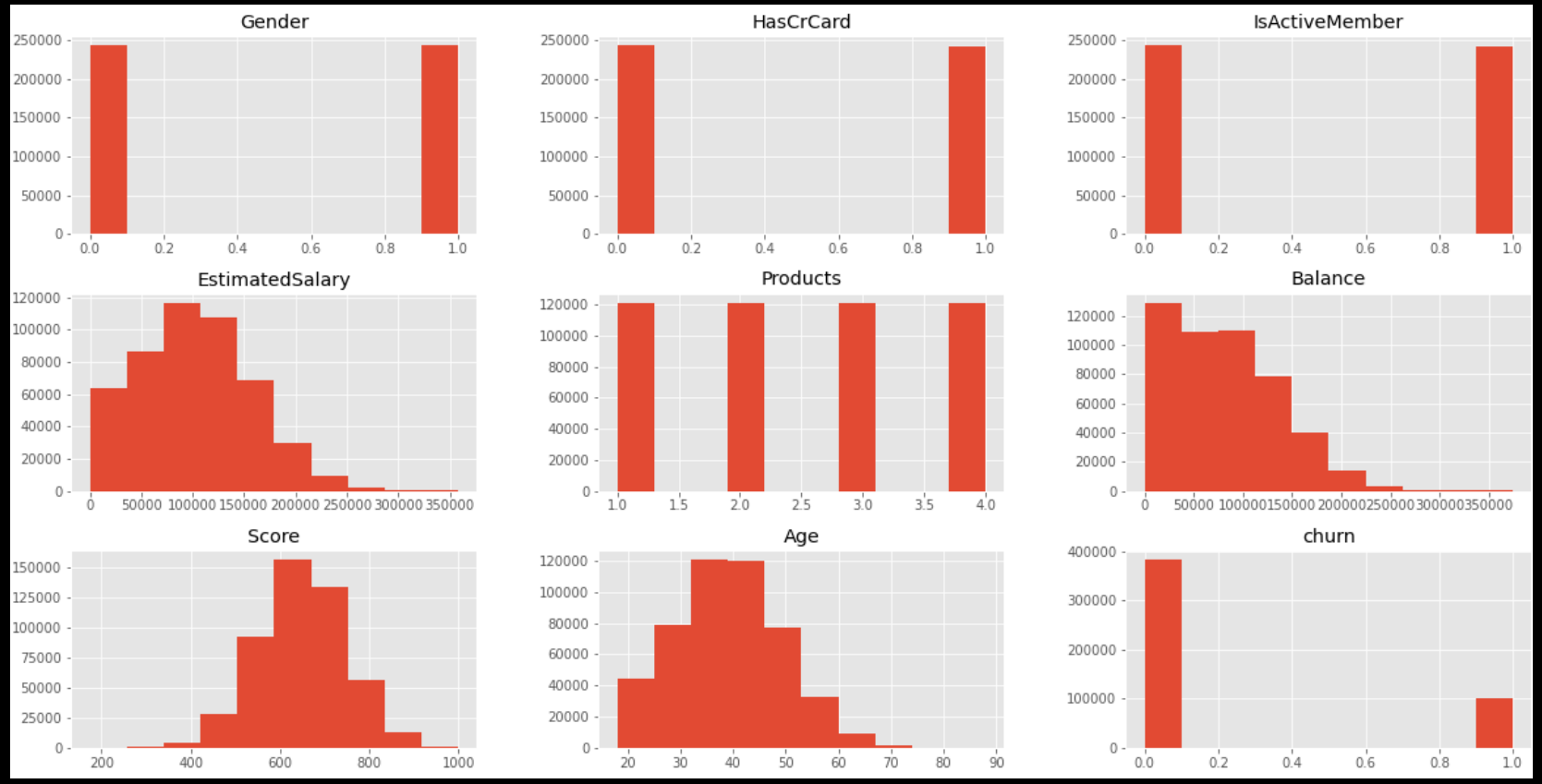
- Products: Número de productos por cliente al momento de aplicar
- Balance: Balance en la cuenta bancaria del cliente al momento de aplicar
- Score: Score crediticio del cliente al momento de aplicar
- Age: Edad del cliente

	Gender	HasCrCard	IsActiveMember	EstimatedSalary	Products	Balance	Score	Age
count	484527.000	484527.000	484527.000	484527.000	484527.000	484527.000	484527.000	484527.000
mean	0.500	0.499	0.499	101124.206	2.499	79760.215	650.166	38.506
std	0.500	0.500	0.500	55520.676	1.118	56652.791	96.726	10.238
min	0.000	0.000	0.000	0.000	1.000	-0.000	174.000	18.000
25%	0.000	0.000	0.000	61332.650	1.000	34261.270	585.000	31.000
50%	0.000	0.000	0.000	100173.440	2.000	76331.900	650.000	38.000
75%	1.000	1.000	1.000	139146.980	3.000	118577.520	715.000	45.000
max	1.000	1.000	1.000	358654.530	4.000	374633.660	1000.000	88.000

Modelado y preprocesamiento

Se debe crear la variable target y codificar las necesarias para que sean aplicables en el modelo

- churn: Variable Target. 1 si el cliente canceló el producto y 0 de otro modo
- Gender: Ahora es binaria. 1 para Male y 0 para Female
- Asimismo, se eliminaron los registros nulos en 'Gender' al no ser imputables y para tratar con los valores nulos. La base resultó con 484527 registros al final.



Elección del Modelo

Para seleccionar el modelo a utilizar, se probó con la línea base de los datos y los hiperparámetros default para conseguir los Accuracy (Scores) de cada algoritmo

Se consideraron 4 algoritmos o clasificadores:

1. KNearestNeighbors
2. Decision Tree (Árbol de decisiones)
3. Random Forest
4. Neural Net (Redes Neuronales)

Como se puede apreciar, los tres últimos algoritmos presentan puntajes aproximadamente iguales en centésimas.

La mejora de accuracy no resulta de gran magnitud, respecto a los otros. Por lo que, al ser más interpretable, se optará por el Decision Tree.

Por otra parte, 484527 registros con 8 variables se pueden considerar como un gran número de data respecto a las variables. Así que, optar por un algoritmo de bajo sesgo y alta varianza como este no resultaría un problema

	models	score
0	Nearest_Neighbors	0.724
1	Decision_Tree	0.792
2	Random_Forest	0.792
3	Neural_Net	0.792

Métricas del Modelo

Finalmente, el modelo usado se le aplicó la técnica de oversampling SMOTE de imblearn para corregir el desbalance de 'churn' y lograr una relación 1:1. Asimismo, se comparan las métricas respecto al modelo de Decision Tree sin aplicar el oversampling.

	accuracy	precision	recall
SMOTE classification	0.5011	0.2079	0.5048
Normal classification (Baseline)	0.7936	0.3	0.0002

Resulta interesante notar que el accuracy sin el oversampling disminuye aproximadamente en 0.26 respecto a la línea base. Sin embargo, considerando el evento sensible (que el cliente cancele el producto) se aprecia una mejora notable en el recall que denotaría que el modelo ahora predice mucho mejor los True Positive (probabilidad de predecir correctamente churn = 1)

Perspectivas del modelo

Juzgando por el nodo principal, se puede concluir que tener una tarjeta de crédito separa de la mejor manera los clientes que cancelan y no el producto antes de los 2 años.

La mayor concentración de pérdidas de clientes se encuentra en las mujeres que no tienen tarjeta de crédito y no son miembros activos.