

# Predicción de pérdida de clientes bancarios

Bank Customer Churn Prediction

## Data Science



Comisión 46295  
Profesor Norberto Leonel González  
Tutor: Gabriel Gutiérrez Mas

## Predicción de pérdida de clientes bancarios2

Introducción .....2

Contexto y problemas comerciales .....3

Definición de Churn..... 3

Contexto y problemas comerciales..... 3

Objetivos e Hipótesis .....4

Objetivo ..... 4

Hipótesis principal..... 4

Desarrollo .....5

Obtención de datos ..... 5

Variables ..... 5

Análisis de datos ..... 5

Desarrollando el modelo ..... 5

Modelos..... 15

## Introducción

El presente trabajo intentará desarrollar un modelo que pueda predecir cuál es la probabilidad de abandono de los clientes de un banco. Con la ayuda del Machine Learning, y con los datos que se presentarán para su análisis, se creará un modelo que buscará patrones que permitan determinar con la mayor certeza la tasa de abandono. De esta manera se podrá ofrecer a los usuarios o interesados una herramienta más que permita tomar decisiones ante determinadas situaciones para prevenir el abandono.



## Definición de Churn

El "Churn de clientes bancarios" se refiere a la tasa de rotación o pérdida de clientes en una entidad bancaria. Es un indicador muy importante y que se puede también aplicar a distintas empresas de servicios. Mide la cantidad de clientes que dejan de utilizar los servicios de un banco en un período de tiempo determinado.

Es relevante para las instituciones financieras, ya que las afecta directamente en su rentabilidad y éxito a largo plazo. Afecta la permanencia de las empresas en determinados mercados. (GPT, s.f.)

## Contexto y problemas comerciales

A diferencia de Argentina, el contexto actual de los bancos en España, Alemania y Francia puede variar, pero en general, los tres países tienen sistemas financieros muy sólidos.

Por hacer una comparación, los bancos en Argentina experimentan situaciones como la alta inflación, el riesgo crediticio o la competencia por captar depósitos de sus clientes.

Es importante destacar que al momento de revisar estos modelos se tendrían que consultar distintas fuentes financieras relacionadas a la industria para poder evaluar cambios como nuevas regulaciones bancarias. (se puede ampliar)

# Objetivos e Hipótesis

## Objetivo

El objetivo principal es desarrollar un modelo predictivo, que permita usando determinadas variables, sugerir medidas para evitar el abandono de los clientes de un banco. (a revisar)

El objetivo secundario es que sirva de referencia para otras empresas de servicios similares. (a revisar)

## Hipótesis principal

“Los que abandonan el banco son los clientes que tienen menor edad y no tienen productos activos como tarjetas de crédito”.

Es decir son clientes inactivos. Se suma a ello, que son mujeres y tienen los salarios más bajos.

Se irán generando preguntas durante el desarrollo que puedan ayudar a resolver el ¿por qué los clientes abandonan el banco de acuerdo a la información que brinda el data set?

## Obtención de datos

La base seleccionada es Churn\_Modelling\_archivo de Kaggle.

El data frame tiene 1000 filas con 14 atributos, de los cuales se determinarán cuáles son los más relevantes.

## Variables

Las variables consideradas por el momento son 'CreditScore', 'Gender', 'Age', 'Tenure', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'EstimatedSalary', 'Exited'.

## Análisis de datos

A partir de esta etapa se comienza el análisis del Df(dataframe) con las variables que se plantearon inicialmente.

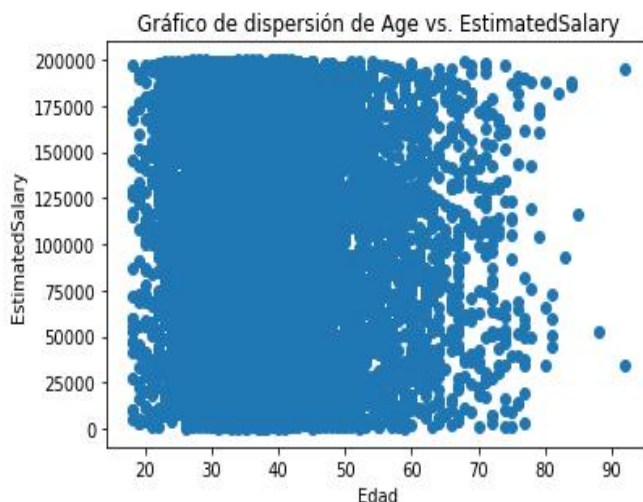
Se irá manipulando el df y generando visualizaciones explicando los avances para la problemática planteada.

## Desarrollando el modelo

El objetivo se dijo es lograr predecir que clientes pueden abandonar un banco.

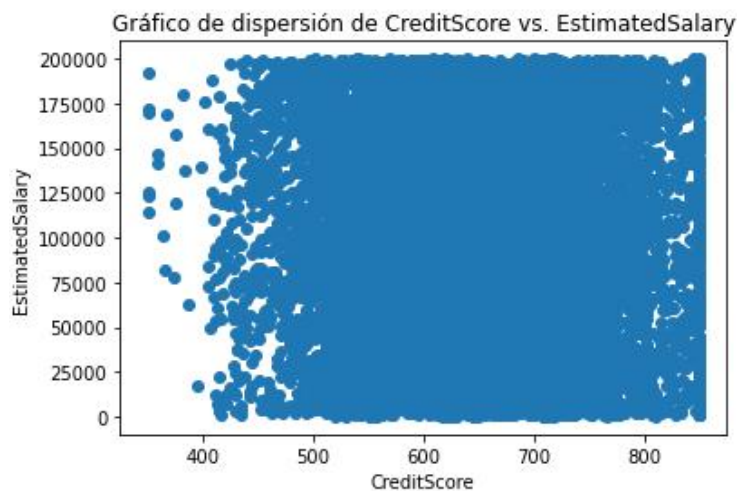
### Edad vs Salario

Se comienza el análisis tratando de relacionar Edad vs Salario con un gráfico con dispersión de cual no puedo sacar muchas conclusiones, salvo que se concentran los salarios estimados entre la edad de 20 hasta los 60 como ocurre en varios países.



## Score vs Salario

Continúa el análisis tratando de relacionar Score Vs Salario. Como se esperaba se pudo afirmar que, a mayor score, mayor salario o viceversa. De todas maneras, se necesita seguir ampliando la información.



## Edad

Si bien es más apropiado cuánto mayor es la cantidad de datos a analizar para obtener conclusiones, por ahora se va a comenzar de menor a mayor para facilitar la interpretación de los datos. Me interesaba conocer las edades de los clientes. Como se vio anteriormente en los gráficos, las edades se concentran entre los 30 y 50 años.

	Edad	Cantidad
0	37	478
1	38	477
2	35	171
3	36	476
4	34	117
5	33	442
6	40	432
7	39	423
8	32	418
9	31	104
10	41	166
11	29	348
12	30	327
13	42	321
14	13	297
15	28	273
16	14	257
17	43	229
18	46	226
19	27	209
20	26	200
21	17	175
22	48	168
23	25	154
24	49	147
25	50	134
26	24	132
27	51	119

También vemos aisladas algunas edades entre 82 y 92 años.



```

Out[5]: 37    478
        38    477
        35    474
        46    474
        34    447
        ...
        92     2
        88     1
        82     1
        84     1
        93     1
        Name: Age, Length: 78, dtype: int64

```

Algunas estadísticas descriptivas sobre el DF para facilitar la lectura de los datos y nos concentramos en la edad, ampliando las conclusiones que se dejaron anteriormente.

```

In [6]: #Algunas estadísticas descriptivas sobre el DataFrame para facilitar la lectura de los datos
        cf_Churn_Modelling_archivo.describe()

```

```

Out[6]:
```

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
count	10000.00000	1.030000e+04	10300.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.539094e+07	550.528300	38.921800	5.012800	73485.889268	1.530200	0.70550	0.515100	100950.239861
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.000000	0.000000	11.580000
25%	2500.75000	1.552853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.000000	0.000000	51002.110000
50%	5000.50000	1.539074e+07	552.000000	37.000000	5.000000	97198.540000	1.000000	1.000000	1.000000	100153.915000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.000000	1.000000	149388.247500
max	10000.00000	1.581569e+07	950.000000	92.000000	10.000000	253898.090000	4.000000	1.000000	1.000000	199962.480000

## Rangos de edad

Decido filtrar por rangos a partir de los resultados del gráfico de dispersión sobre Edad vs Score. En este caso entre 20 a 59 que concentra la mayor cantidad:

Edad	Cantidad
0	37
1	38
2	35
3	36
4	34
5	33
6	40
7	39
8	32
9	31
10	41
11	29
12	30
13	42
14	43
15	28
16	44
17	45
18	46
19	27



20	26	200
21	47	175
22	48	168
23	25	154
24	49	147
25	50	134
26	24	132
27	51	119
28	52	102
29	23	99
30	54	84
31	22	84
32	55	82
33	57	75
34	53	74
35	56	70
36	58	67
37	59	62
38	21	53
39	20	40

Filtro por rango de edad para que solo traiga los de 30 ya que seguía siendo mucha información.

RowNumber	CustomerId	Surname	Creditscore	Geography	Gender	Age
3	4	Boni	699	France	Female	39
10	11	Beance	528	France	Male	31
12	13	Kay	476	France	Female	34
14	15	Scott	635	Spain	Female	35
21	22	Dellucci	636	Spain	Female	32
...	...	...	...	...	...	...
9990	9991	Nkemakonam	714	Germany	Male	33
9992	9993	Chukwualuka	726	Spain	Male	36
9995	9996	Obijiaku	771	France	Male	39
9996	9997	Johnstone	516	France	Male	35
9997	9998	Liu	709	France	Female	36

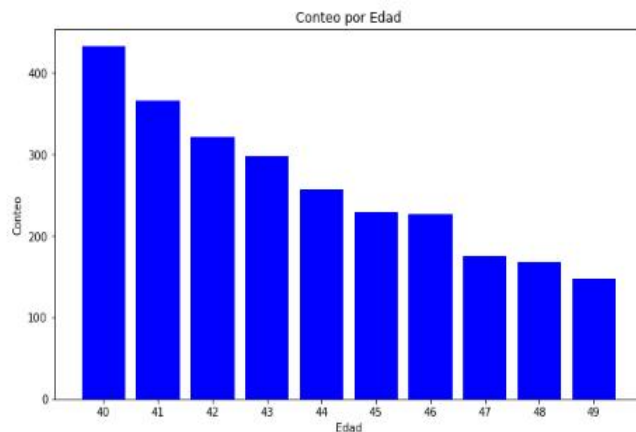
Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	\
3	1	0.00	2	0	0
10	6	102016.72	2	0	0
12	10	0.00	2	1	0
14	7	0.00	2	1	1
21	8	0.00	2	1	0
...	...	...	...	...	...
9990	3	35016.60	1	1	0
9992	2	0.00	1	1	0
9995	5	0.00	2	1	0
9996	10	57369.61	1	1	1
9997	7	0.00	1	0	1

EstimatedSalary	Exited	
3	93826.63	0
10	80181.12	0
12	26260.98	0
14	65951.65	0
21	138555.46	0
...	...	...
9990	53667.08	0
9992	195192.40	0
9995	96270.64	0
9996	101699.77	0
9997	42085.58	1

[4346 rows x 14 columns]

Luego se filtra por los de 40 años y se grafica como para ir teniendo un panorama:

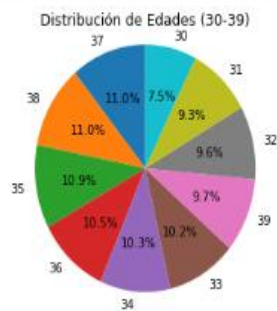


Ahora vuelvo a los de 30 años buscando más adelante ir haciendo una comparación:

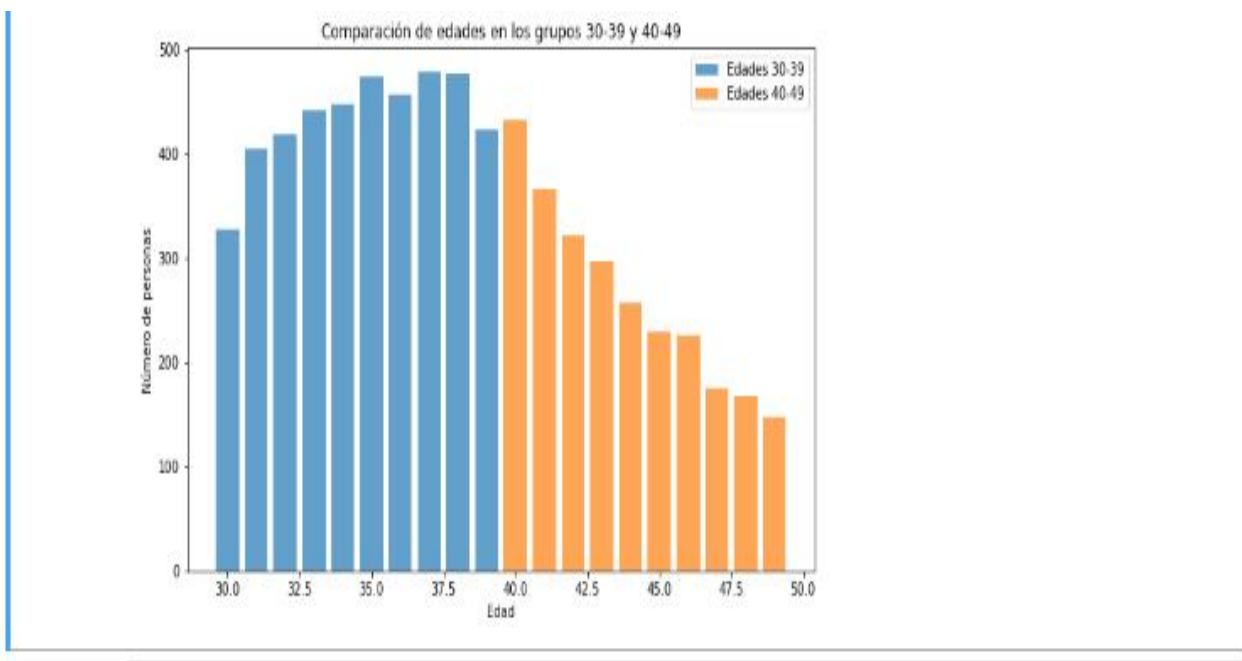
```
In [8]: # Filtro por Los de 30s para contar La cantidad que se repite de cada edad pero siempre con Los de 30
Age_df_30 = df_Churn_Modelling_archivo[(df_Churn_Modelling_archivo['Age'] >= 30) & (df_Churn_Modelling_archivo['Age'] <= 39)]

# Contar Las ocurrencias de cada categoría
age_counts = Age_df_30['Age'].value_counts()

# Crear el gráfico de torta con Matplotlib
plt.pie(age_counts, labels=age_counts.index, autopct='%1.1f%%', startangle=90)
plt.axis('equal') # Hace que el gráfico sea un círculo
plt.title('Distribución de Edades (30-39)')
plt.show()
```



Ahora puedo comparar los de 30(azul) años y 40 años(naranja):

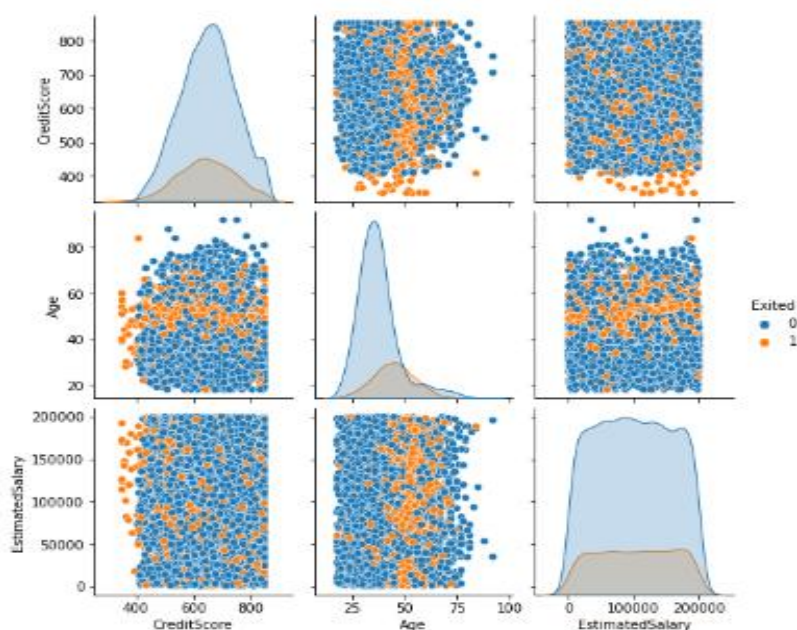


La conclusión más importante que obtengo de los análisis de las edades es que a medida que aumenta la edad bajan las cantidades de personas o clientes del banco.

Ahora uso seaborn para mostrar lo que consideré más importante buscando relaciones, esto es Credit Score, Edad, etc

```
In [15]: #Usa seaborn para mostrar lo que considere más importante buscando relaciones, esto es Credit Score, Edad etc
importantes = df_Churn_Modelling_archivo[['CreditScore', 'Age', 'EstimatedSalary', 'Exited']]
sns.pairplot(importantes, hue = 'Exited')

Out[15]: <seaborn.axisgrid.PairGrid at 0x24ea89bb190>
```



Como se puede observar, las posibilidades que brinda esta visualización es bastante amplia. Me permite entre otras cosas, inferir rápidamente cuáles son los clientes que abandonan el banco (Exited = 1), por Creditscore, por Edad(Age) y por EstimatedSalary. Por ejemplo, observo que los que tienen menor CreditScore son los que abandonan el banco, o los que tienen menor EstimatedSalary.

Como se vio anteriormente, la edad no sería un factor determinante de abandono del banco. Por lo cual, voy en busca de la/las variables.

## Género

Necesito ver si el género influye, por lo cual lo verifico:

```
In [10]: #Acá vamos a ver ahora que cantidades tenemos de hombres y mujeres
df_Churn_Modelling_archivo.describe(include=['O'])
```

```
Out[10]:
```

	Surname	Geography	Gender
count	10000	10000	10000
unique	2932	3	2
top	Smith	France	Male
freq	32	5014	5457

```
In [ ]: #Se observa que es mayor la cantidad de hombres, pero no sabemos en que cantidad
```

Se verifica que en el top están los hombres, pero no sabemos si es grande la diferencia por sobre las mujeres, así que se sigue desarrollando.

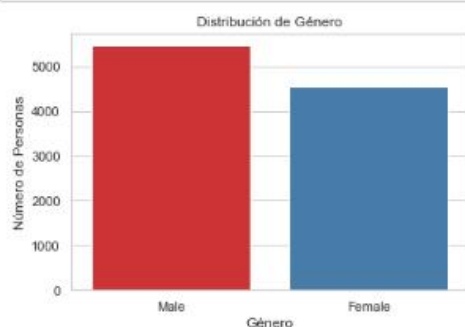
Se grafica por o cuál Mujeres vs Hombres:

```
In [77]: # Por último me interesaba saber del set la distribución de los géneros, por lo cual es solo contar en este caso los M y F,
gender_counts = df_Churn_Modelling_archivo['Gender'].value_counts()

sns.barplot(x=gender_counts.index, y=gender_counts.values, palette="Set1")

plt.xlabel('Género')
plt.ylabel('Número de Personas')
plt.title('Distribución de Género')

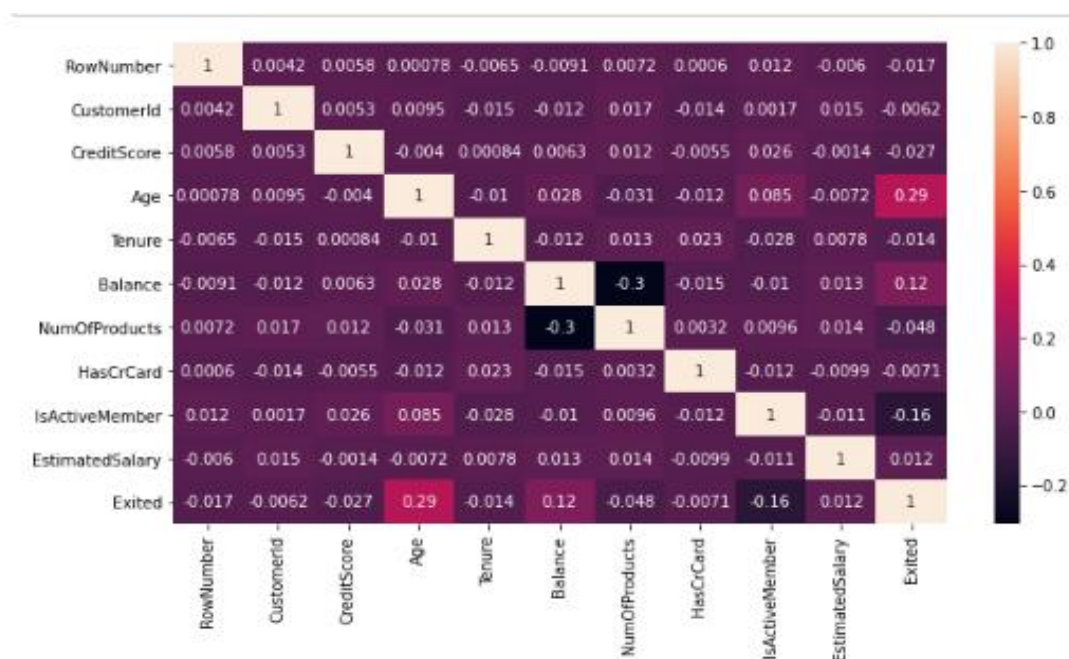
plt.show()
```



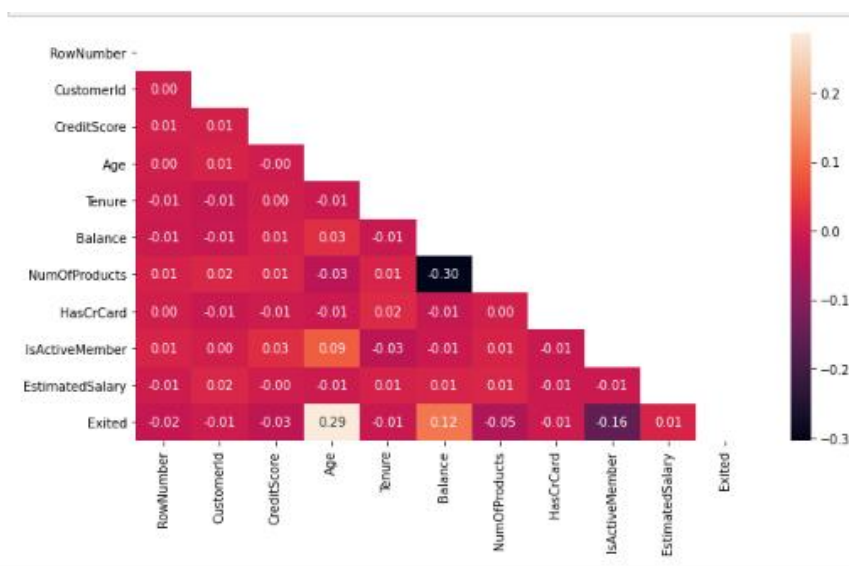
Vemos que la distribución Male vs Female es muy parecida, casi un 50% de c/u. Por lo cual, puedo inferir que el sexo no sería un determinante.

## Análisis de correlaciones

Considero que, para poder seguir un camino aceptable para conseguir el modelo, se tendría que hacer un análisis de correlaciones.



Para seguir aclarando lo anterior, se descartan las variables que no son relevantes, resultando lo siguiente:





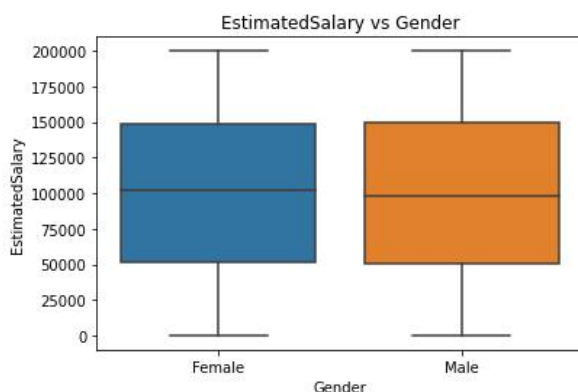
Se observa que la edad, o el género, no son relevantes. Pero si aparece una variable que podría guiarnos para lograr lo que deseamos del modelo. Esto es El número de productos que tendría un cliente NumOfProducts. Continuaremos el análisis por esa variable más adelante.

### EstimatedSalary distribuidos por Gender

Siguiendo con el análisis de EstimatedSalary distribuidos por Gender, veo que sigue siendo 50 y 50.

```
In [ ]: #Siguiendo con el análisis de EstimatedSalary distribuidos por Gender, veo que sigue siendo 50 y 50.
```

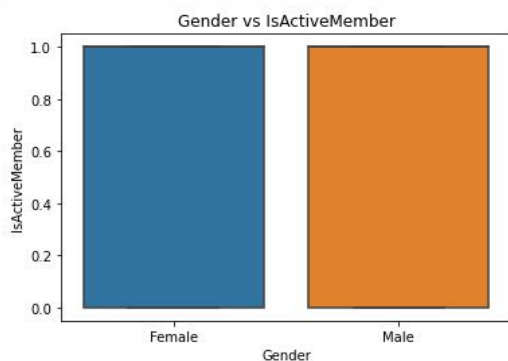
```
In [20]: sns.boxplot(x='Gender', y = 'EstimatedSalary', data = df_Churn_Modelling_archivo)
#Data.boxplot(grid= False, column = ['pay'], by = ['gender'])
plt.title("Gender vs EstimatedSalary");
```



### Género de los clientes activos

Por la relación anterior, tenía que ser 50 y 50.

```
In [24]: # Quiero saber sexos de los clientes activos
#Por la relación anterior, tenía que ser 50 y 50.
sns.boxplot(x='Gender', y = 'IsActiveMember', data = df_Churn_Modelling_archivo)
#Data.boxplot(grid= False, column = ['pay'], by = ['gender'])
plt.title("Gender vs IsActiveMember");
```





## Cientes Activos vs Inactivos

Se vio que la edad posiblemente, no es un factor determinante para que los clientes abandonen el banco.

Por ello me interesa saber la actividad de los clientes. Por acá pienso que se puede obtener alguna respuesta, para saber quiénes abandonan o no el banco:

a- Cuento clientes que están activos:

```
df_Churn_Modelling_archivo['IsActiveMember'].value_counts()
1      7055(activos)
0      2945(inactivos)
Name: HasCrCard, dtype: int64
```

b- Cuento los clientes que tienen tarjetas de Crédito, por lo cual siguen activos:

```
df_Churn_Modelling_archivo['HasCrCard'].value_counts()
1      7055
0      2945
Name: HasCrCard, dtype: int64
```

c- Cuento clientes que se fueron vs los que siguen, buscando respuestas a las hipótesis:

```
df_Churn_Modelling_archivo['Exited'].value_counts()
0      7963(activos)
1      2037(inactivos)
Name: Exited, dtype: int64
```

## Modelos

### Modelos I- EstimatedSalary - Age

Volviendo al tema edad, quiero averiguar qué resultados obtendría con un modelo para predecir el EstimatedSalary según la Edad. Esto es, saber si fue correcto que afirmé que no había correlación entre la edad y las demás variables que determinen el abandono. Se presenta lo siguiente:

```
model1 = 'EstimatedSalary~Age'
```

OLS Regression Results						
Dep. Variable:	EstimatedSalary	R-squared:	0.000			
Model:	OLS	Adj. R-squared:	-0.000			
Method:	Least Squares	F-statistic:	0.01916			
Date:	Tue, 24 Oct 2023	Prob (F-statistic):	0.890			
Time:	00:56:17	Log-Likelihood:	-1.2379e+05			
No. Observations:	10000	AIC:	2.476e+05			
Df Residuals:	9998	BIC:	2.476e+05			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	1.006e+05	3913.640	25.712	0.000	9.3e+04	1.08e+05
CreditScore	-0.8237	5.951	-0.138	0.890	-12.488	10.841
Omnibus:	7392.705	Durbin-Watson:	2.010			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	581.607			
Skew:	0.002	Prob(JB):	5.08e-127			
Kurtosis:	1.819	Cond. No.	4.48e+03			

Se puede ver que el Pv es mayor a 0,05 por lo cual vamos a descartar la variable edad nuevamente.

## Modelo 2- EstimatedSalary - CreditScore

model2 = 'EstimatedSalary~CreditScore'

OLS Regression Results						
Dep. Variable:	EstimatedSalary	R-squared:	0.000			
Model:	OLS	Adj. R-squared:	-0.000			
Method:	Least Squares	F-statistic:	0.01916			
Date:	Tue, 24 Oct 2023	Prob (F-statistic):	0.890			
Time:	00:56:17	Log-Likelihood:	-1.2379e+05			
No. Observations:	10000	AIC:	2.476e+05			
Df Residuals:	9998	BIC:	2.476e+05			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	1.006e+05	3913.640	25.712	0.000	9.3e+04	1.08e+05
CreditScore	-0.8237	5.951	-0.138	0.890	-12.488	10.841
Omnibus:	7392.705	Durbin-Watson:	2.010			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	581.607			
Skew:	0.002	Prob(JB):	5.08e-127			
Kurtosis:	1.819	Cond. No.	4.48e+03			

Se puede ver que el Pv es mayor a 0,05 por lo cual vamos a descartar la variable CreditScore.

### **Modelo 3 - IsActiveMember - EstimatedSalary - Age - NumOfProducts**

Pruebo con otras variables para el modelo con el IsActiveMember segun El salario estimado, la edad y el número de productos.

model3 = IsActiveMember~EstimatedSalary - Age - NumOfProducts

OLS Regression Results						
=====						
Dep. Variable:	IsActiveMember	R-squared:	0.008			
Model:	OLS	Adj. R-squared:	0.007			
Method:	Least Squares	F-statistic:	25.44			
Date:	Wed, 22 Nov 2023	Prob (F-statistic):	2.19e-16			
Time:	20:23:28	Log-Likelihood:	-7215.3			
No. Observations:	10000	AIC:	1.444e+04			
Df Residuals:	9996	BIC:	1.447e+04			
Df Model:	3					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
Intercept	0.3493	0.025	13.942	0.000	0.300	0.398
EstimatedSalary	-9.542e-08	8.66e-08	-1.102	0.271	-2.65e-07	7.43e-08
Age	0.0041	0.000	8.604	0.000	0.003	0.005
NumOfProducts	0.0107	0.009	1.244	0.214	-0.006	0.027
=====						
Omnibus:	35041.681	Durbin-Watson:	2.015			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1618.984			
Skew:	-0.054	Prob(JB):	0.00			
Kurtosis:	1.032	Cond. No.	5.91e+05			

Puedo ver que mejoró el R-squared con respecto a los modelos anteriores, pero tendría que ir probando con más variables, agregando o sacando etc. Esto se va a ver más a delante desarrollando los temas.