Predicción de pérdida de clientes bancarios

Bank Customer Churn Prediction

Data Science

Comisión 46295
Profesor Norberto Leonel González
Tutor: Gabriel Gutiérrez Mas

CONTENIDO

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 Learnig, 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.

Contexto y problemas comerciales

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?

Desarrollo

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 Originales

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

Variables Modificadas

Se modificaron aquellas variables que era de texto a número para facilitar el análisis quedando las siguientes:

```
Index(['RowNumber', 'CustomerId', 'Surname', 'CreditScore', 'Geography',
       'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard',
       'IsActiveMember', 'EstimatedSalary', 'Exited', 'Gender_modificado',
       'Geography_modificado'],
      dtype='object')
```

Análisis de datos

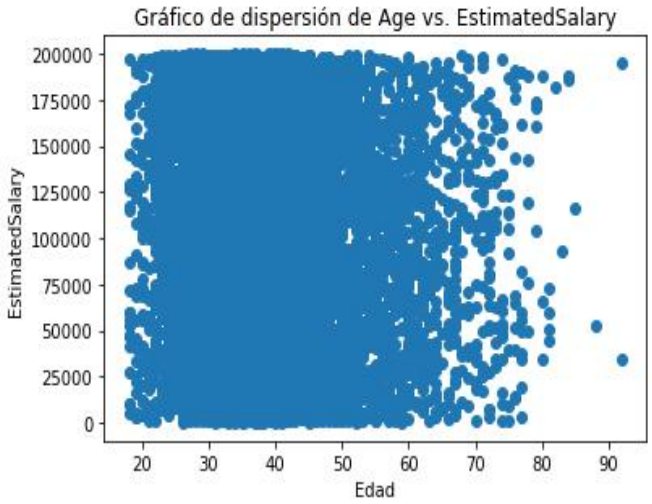
A partir de esta etapa se comienza el análisis del Df(dataframe) con las variables que se plantearon inicialmente.
Se irá manipulando el df y generando visualizaciones explicando los avances para la problemática planteada.

Desarrollando el modelo

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

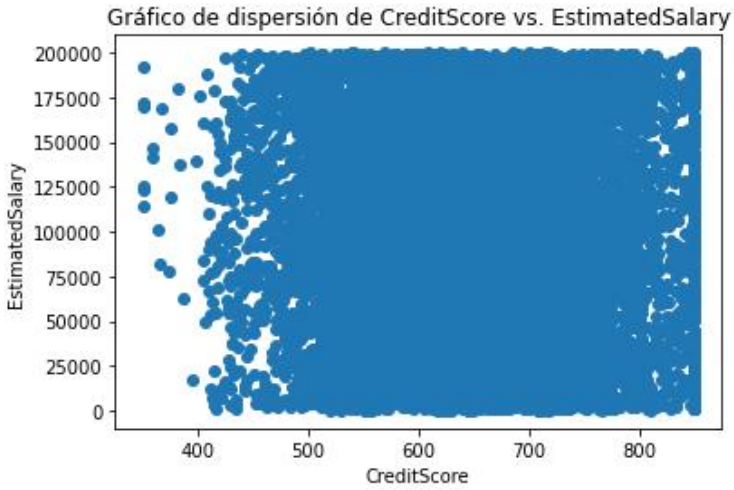
Edad vs Salario

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 pude 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	474
3	36	456
4	34	447
5	33	442
6	40	432
7	39	423
8	32	418
9	31	404
10	41	366
11	29	348
12	30	327
13	42	321
14	43	297
15	28	273
16	44	257
17	45	229
18	46	226
19	27	209
20	26	200
21	47	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
        36    456
        34    447
              ...
        92      2
        88      1
        82      1
        85      1
        83      1
        Name: Age, Length: 70, 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 algunas estadísticas descriptivas sobre el DataFrame para facilitar la lectura de los datos
        df_Churn_Modelling_archivo.describe()
```

Out[6]:

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881
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	0.00000	0.000000	11.580000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000

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	478
1	38	477

2	35	474
3	36	456
4	34	447
5	33	442
6	40	432
7	39	423
8	32	418
9	31	404
10	41	366
11	29	348
12	30	327
13	42	321
14	43	297
15	28	273
16	44	257
17	45	229
18	46	226
19	27	209
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.

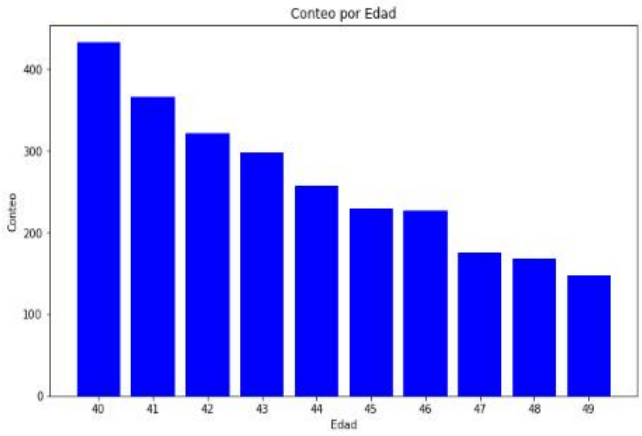
	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age
3	4	15701354	Boni	699	France	Female	39
10	11	15767821	Bearce	528	France	Male	31
12	13	15632264	Kay	476	France	Female	34
14	15	15600882	Scott	635	Spain	Female	35
21	22	15597945	Dellucci	636	Spain	Female	32
...	...	...	...	...	...	...	...
9990	9991	15798964	Nkemakonam	714	Germany	Male	33
9992	9993	15657105	Chukwualuka	726	Spain	Male	36
9995	9996	15606229	Obijiaku	771	France	Male	39
9996	9997	15569892	Johnstone	516	France	Male	35
9997	9998	15584532	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:

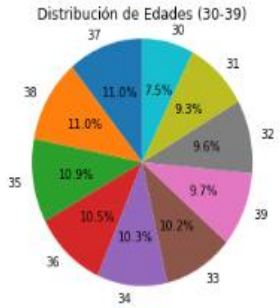


Ahora vuelvo a los de 30 años buscando más adelante ir haciendo una comparación:

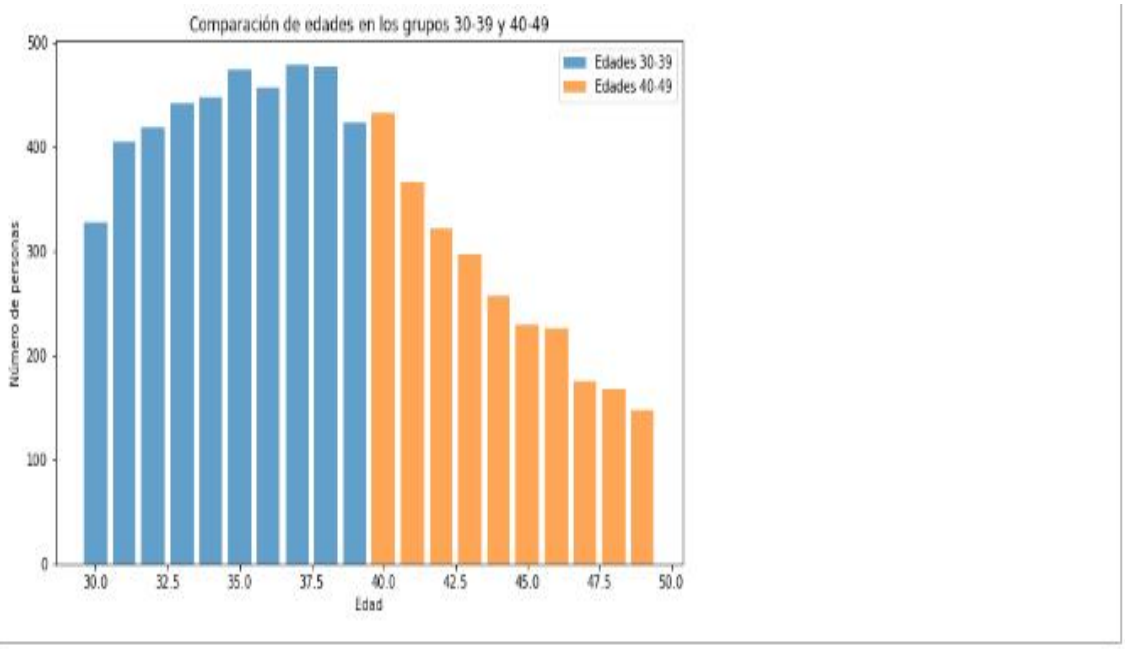
```
In [8]: # Filtro por los de 30s para contar la cantidad que se repite de cada edad pero siempre con los de 30
Age_df_30 = df_Churn_Modelling_archivo[(df_Churn_Modelling_archivo['Age'] >= 30) & (df_Churn_Modelling_archivo['Age'] <= 39)]

# 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):



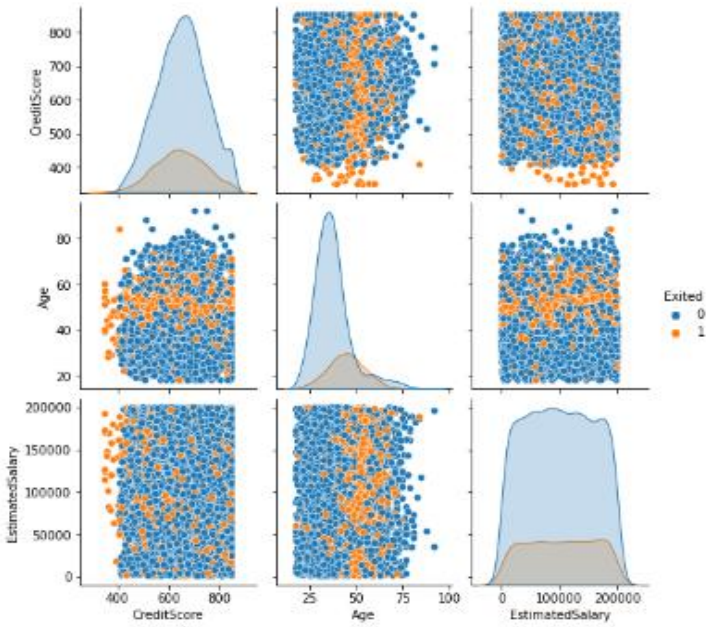
La conclusión más importante que obtengo de los análisis de las edades es que a medida que aumenta la edad bajan las cantidades de personas o clientes del banco.

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

```
In [15]: #Uso seaborn para mostrar lo que consiré 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 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:

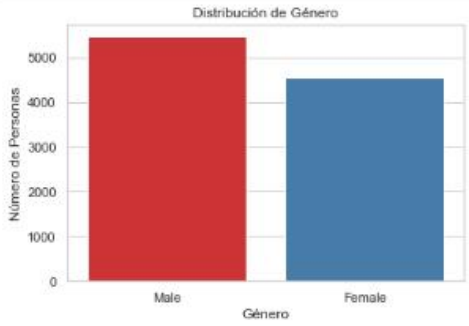
```
In [77]: # Por último me interesaba saber del set la distribución de los géneros, por lo cual es solo contar en este caso los M y F,

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()
```



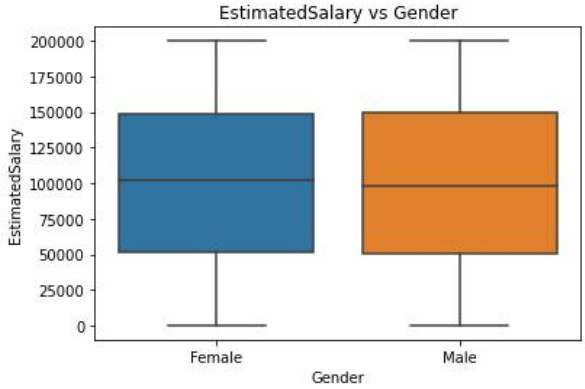
Vemos que la distribución Male vs Female es muy parecida, casi un 50% de c/u. Por lo cual, puedo inferir que el sexo no sería un determinante.

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.
```

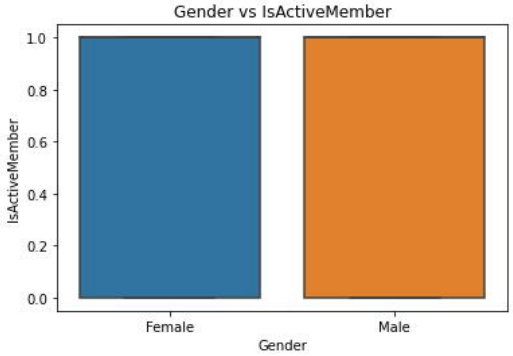
```
In [20]: sns.boxplot(x='Gender', y = 'EstimatedSalary', data = df_Churn_Modelling_archivo)
         #Data.boxplot(grid= False, column = ['pay'], by = ['gender'])
         plt.title("Gender vs EstimatedSalary");
```



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");
```



Clientes 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:

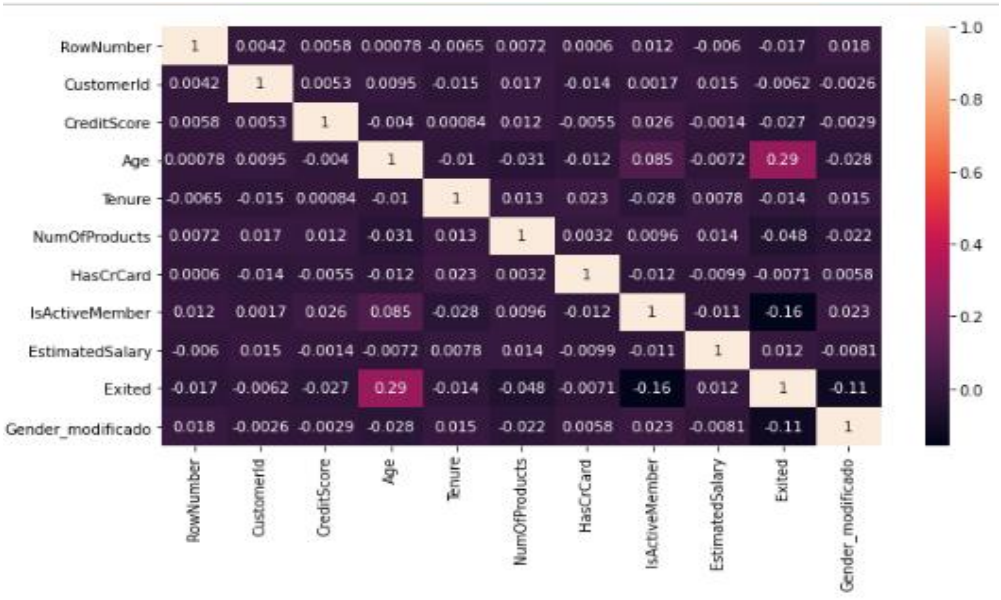
```
df_Churn_Modelling_archivo['HasCrCard'].value_counts()
1    7055
0    2945
Name: HasCrCard, dtype: int64
```

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

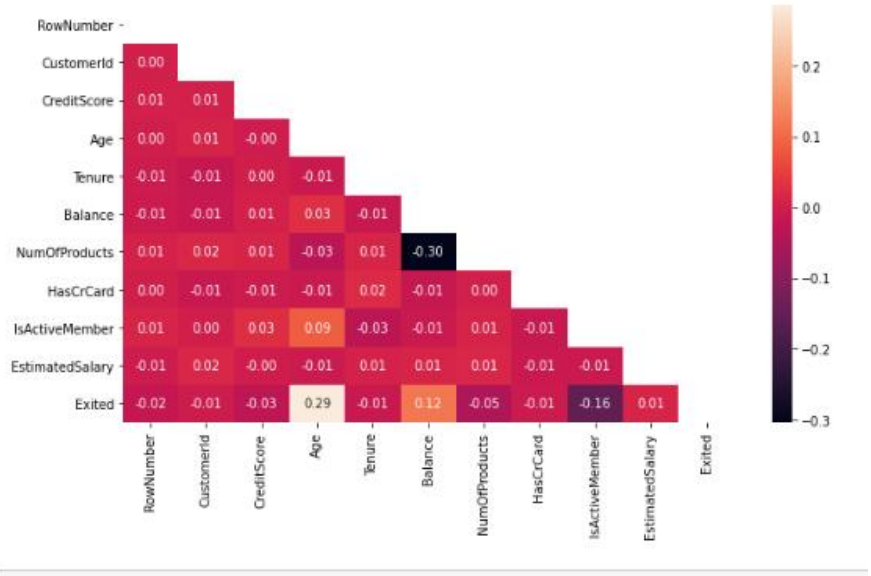
```
df_Churn_Modelling_archivo['Exited'].value_counts()
0    7963(activos)
1    2037(inactivos)
Name: Exited, dtype: int64
```

Análisis de correlaciones

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



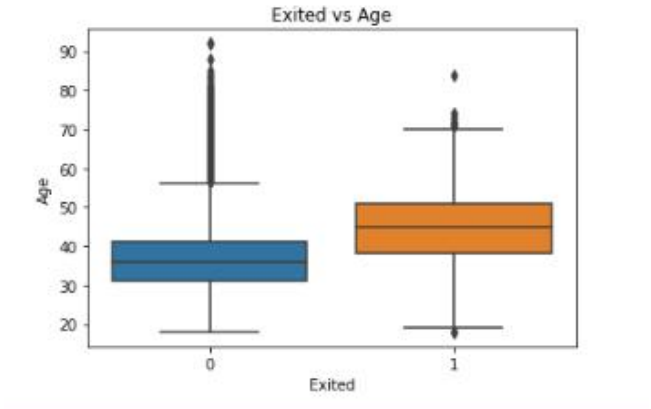
Para seguir aclarando lo anterior, se descartan las variables que no son relevantes, resultando lo siguiente:



Se observa que la relación más importante se daría entre en Exited en función de Age.

Modelos

Previo a armar los modelos quiero veras las cantidades de Exited y la relación con la Age.
Observo los Exited distribuidos por Age, veo que sigue siendo 50 y 50.



Ahora voy a contar los clientes que se fueron, entiendo que serían de la columna Exited = 1, y los que siguen = 0

```
In [51]: #Ahora voy a contar los clientes que se fueron, entiendo que serían de la columna Exited = 1, y los que siguen = 0

         #Cuento clientes que se fueron vs los que siguen , posiblemente por acá pueda  obtener más respuestas a las hipótesis , para sabe
         df_Churn_Modelling_archivo['Exited'].value_counts()

Out[51]: 0    7963
         1    2037
         Name: Exited, dtype: int64
```

Modelos 1- Exited - Age

Busco predecir la probabilidad de salir (Exited) en función de la edad con un modelo con la variable dependiente Exited y como variable independiente a Age. Surge del gráfico de correlación en donde se observó esa relación. Se presenta lo siguiente:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Exited   R-squared:                       0.081
Model:                            OLS   Adj. R-squared:                  0.081
Method:                 Least Squares   F-statistic:                     886.1
Date:                Thu, 18 Jan 2024   Prob (F-statistic):          1.24e-186
Time:                        16:22:58   Log-Likelihood:                -4670.4
No. Observations:               10000   AIC:                             9345.
Df Residuals:                    9998   BIC:                             9359.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.2228      0.015    -15.014      0.000      -0.252      -0.194
Age            0.0110      0.000     29.767      0.000       0.010       0.012
==============================================================================
Omnibus:                     1669.435   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2669.513
Skew:                           1.262   Prob(JB):                         0.00
Kurtosis:                       3.183   Cond. No.                         155.
==============================================================================
```

Que se concluye:

1. R-cuadrado (R-squared): El R-cuadrado es 0.081, lo que significa que alrededor del 8.1% de la variabilidad en la variable dependiente (Exited) puede explicarse por la variable independiente (Age).
2. P>|t|: Ambos valores p son muy cercanos a cero (0.000), lo que sugiere que ambas variables son estadísticamente significativas.
3. F-statistic: El valor es 886.1, y la probabilidad asociada también es 1.24e-186. Este F-statistic y su probabilidad evalúan la significancia global del modelo.

La variable "Age" tiene un coeficiente positivo, lo que sugiere que hay una relación positiva entre la edad y la probabilidad de salir como se viene afirmando.

Pero necesito mejorar el R- R-squared por lo cual voy a intentar agregando al modelo otro coeficiente como "IsActiveMember".Para ello se instaló Statsmodels.

Modelos 11- Exited – Age - IsActiveMember

Ahora sabiendo que es probable que la edad pueda que ser un factor determinante, me interesa saber la actividad de los clientes:
#Cuento clientes activos vs inactivos, posiblemente por acá pueda obtener alguna respuesta, para saber quiénes abandonan o no el banco
df_Churn_Modelling_archivo['IsActiveMember'].value_counts()

```
Out[25]: 1    5151
         0    4849
         Name: IsActiveMember, dtype: int64
```

Entonces ahora Busco predecir la probabilidad de salir (Exited) en función de la edad con un mbdelo con la variable dependiente Exited y como variable independiente a Age y agrego IsactiveMember:

```
# Asumo que df_Churn_Modelling_archivo es tu DataFrame

# Modelo con Exited como variable dependiente y Age y IsActiveMember como variables independientes
model3 = 'Exited ~ Age + IsActiveMember'

# Utilizamos la función ols() para especificar el modelo de regresión lineal
lm3 = smf.ols(formula=model3, data=df_Churn_Modelling_archivo).fit()

# Imprimimos el resumen del modelo
print(lm3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Exited   R-squared:                       0.114
Model:                            OLS   Adj. R-squared:                  0.114
Method:                 Least Squares   F-statistic:                     644.6
Date:                Thu, 18 Jan 2024   Prob (F-statistic):          4.70e-264
Time:                        16:51:03   Log-Likelihood:                -4488.4
No. Observations:               10000   AIC:                             8983.
Df Residuals:                    9997   BIC:                             9005.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1705      0.015    -11.505      0.000      -0.200      -0.141
Age            0.0116      0.000     31.846      0.000       0.011       0.012
IsActiveMember    -0.1465      0.008    -19.248      0.000      -0.161      -0.132
==============================================================================
Omnibus:                     1606.552   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2523.129
Skew:                           1.226   Prob(JB):                         0.00
Kurtosis:                       3.201   Cond. No.                         159.
==============================================================================

Notes:
```

1. R-cuadrado (R-squared): El R-cuadrado es 0.114. Esto significa que alrededor del 11.4% de la variabilidad en la variable dependiente "Exited" puede explicarse por las variables independientes "Age" e "IsActiveMember".

Aunque no es muy alto, sugiere que el modelo explica un cierto porcentaje de la variabilidad.

2. Valores p (P>|t|): Todos los valores p asociados con los coeficientes son muy cercanos a cero (0.000), lo que sugiere que todas las variables son estadísticamente significativas.
3. F-statistic: El valor del estadístico F es 644.6, y su probabilidad asociada es muy cercana a cero (4.70e-264). Esto indica que al menos una de las variables independientes tiene un efecto significativo en la variable dependiente.

En resumen, el modelo sugiere que tanto la edad ("Age") como la membresía activa ("IsActiveMember") están asociadas con la probabilidad de salir ("Exited"). Sin embargo, el R-cuadrado indica que solo el 11.4% de la variabilidad de la variable dependiente se explica con estas dos variables en el modelo, pero subió un 3% aproximadamente con respecto al Modelo I.