Caso de Uso: Fuga de clientes

Equipo Nro 5

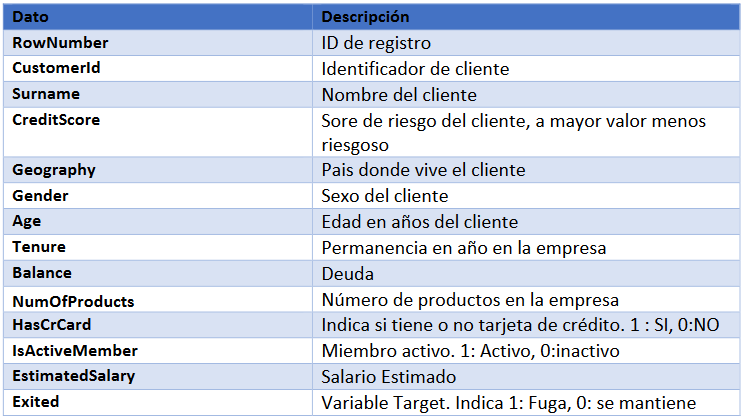
Integrantes: Jordan Rodriguez, José Carlos, Leslie, Andres,

# Objetivo:

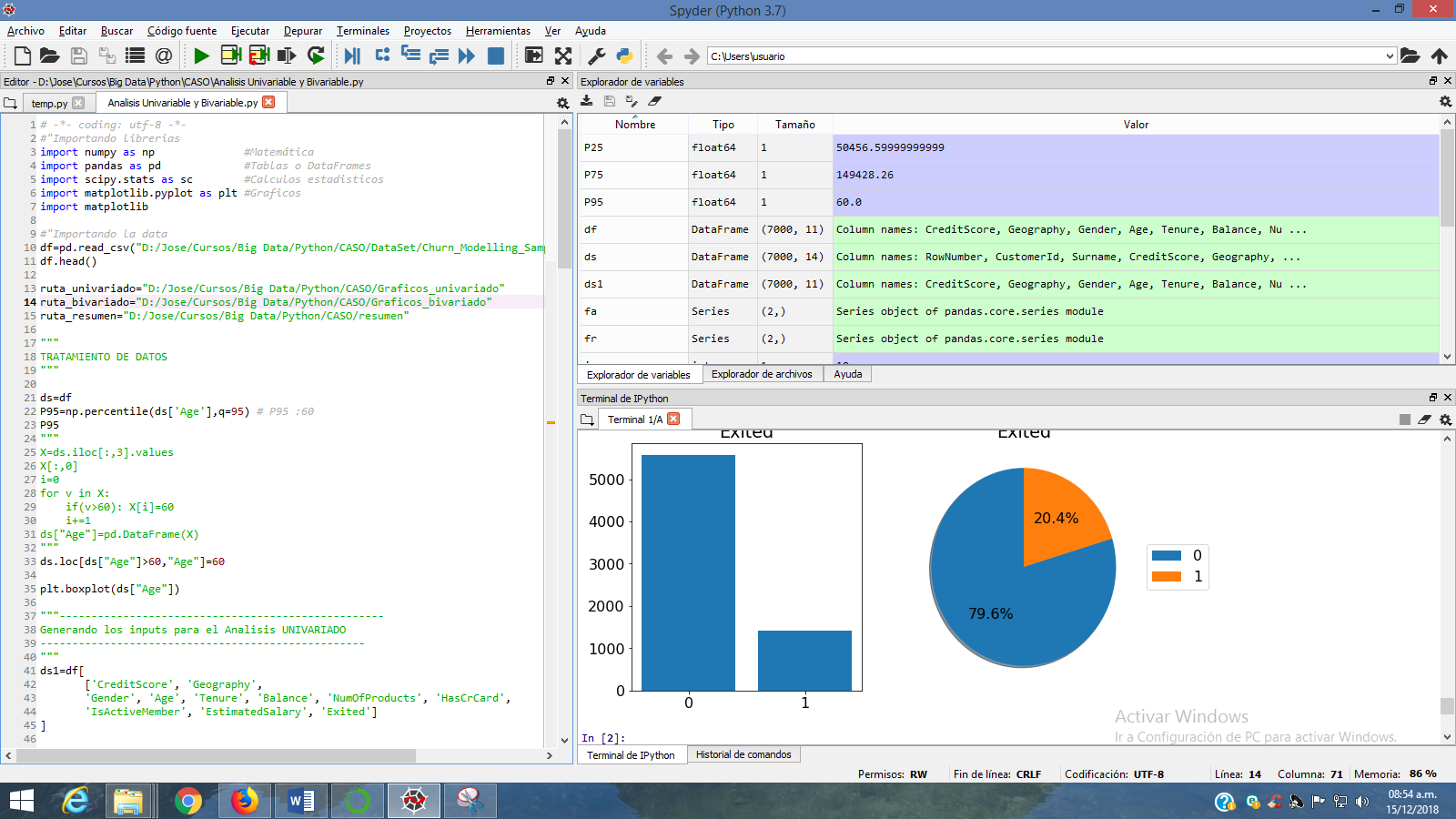
## Predecir la fuga de clientes

Como parte de la gestión de clientes (CRM) es muy común saber quiénes son los clientes que probablemente nos abandonen, de tal manera que se puede tomar acciones para retenerlos o simplemente dejarlos ir. El presente caso de uso consiste en desarrollar el mejor modelo posible, producto de la selección y tratamiento adecuado de las variables, así como la elección del mejor algoritmo de clasificación que permita una validación exitosa a través de la Matriz de Confusión.

### El diseño de registro de la base es la siguiente:



# Análisis Univariado:

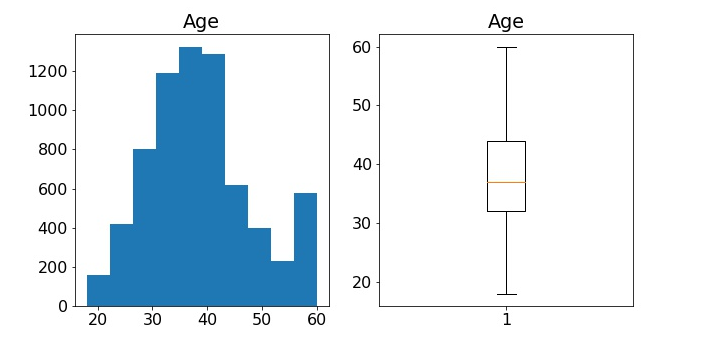


## Variables No Categóricas:

#### Edad:



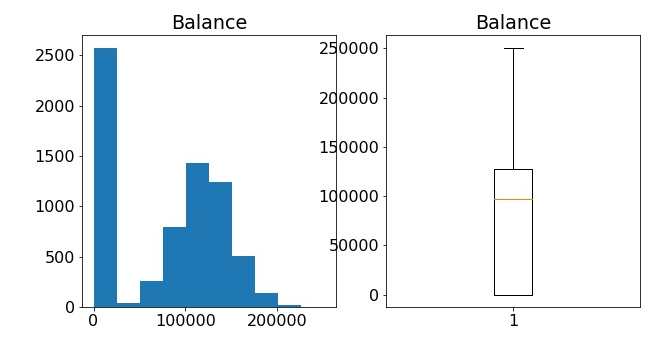
Es una variable distribuida, con Curtosis negativo por lo cual los valores tienen tendencia hacia valores menores que la mediana. No se tiene valores perdidos o nulos.



#### Saldo:



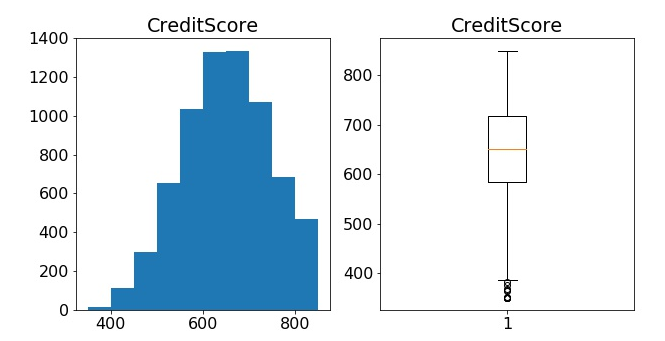
Es una variable que necesita ser ajustada, porque tine muchos valores en el primer quartil y con valores cero.



#### Score Crediticio:

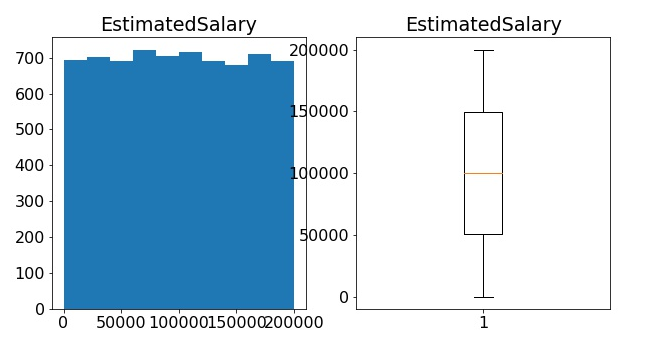


Es una variable distribuida, pero con valores atípicos.



#### Salario:

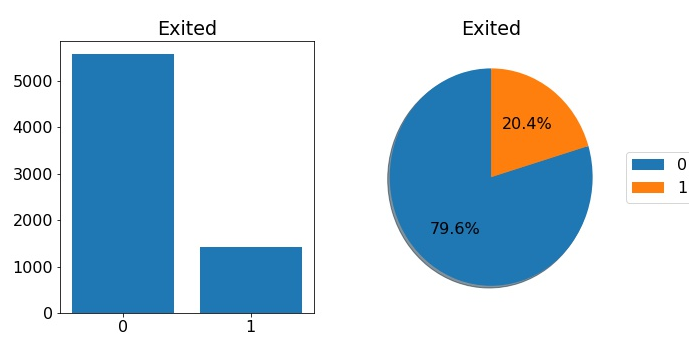




## Variables Categóricas:

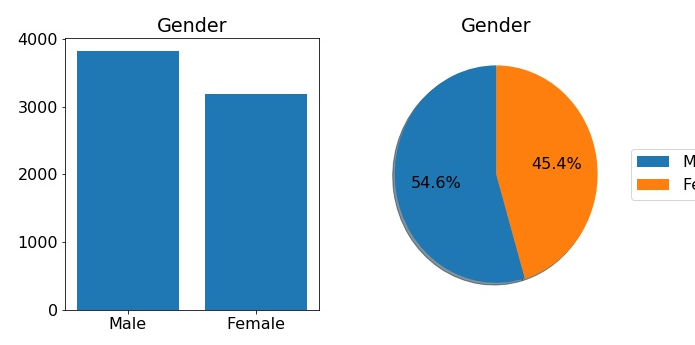
### Bajas:

El 20% de los clientes terminan dando de baja a sus productos.



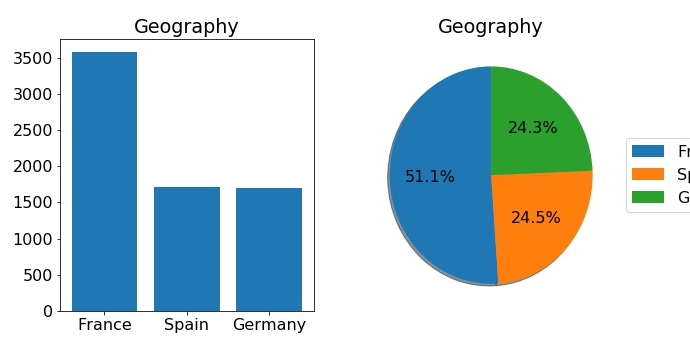
### Género:

El 55% de los clientes son de género masculino.



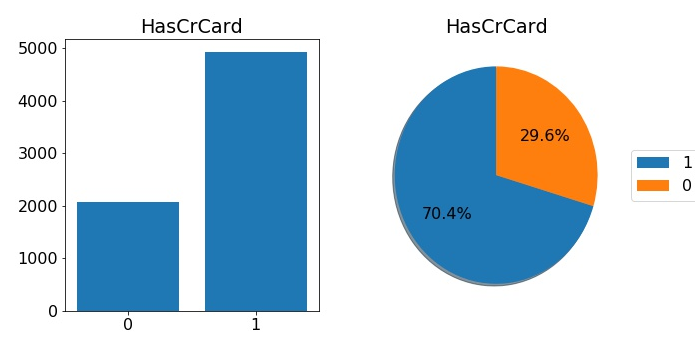
### Ubicación Geográfica:

El 51% de los clientes son de Francia.



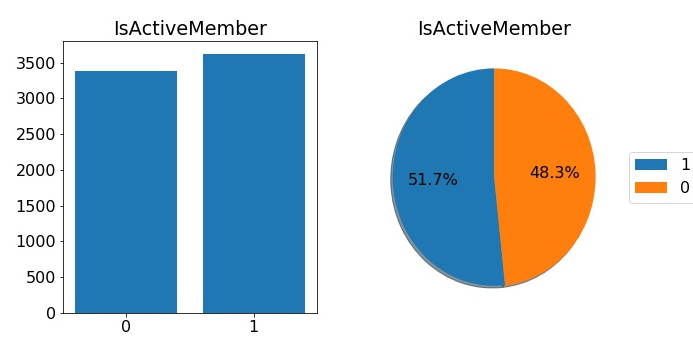
### Tiene Tarjeta de Crédito:

El 70% de clientes tienen tarjeta de crédito.



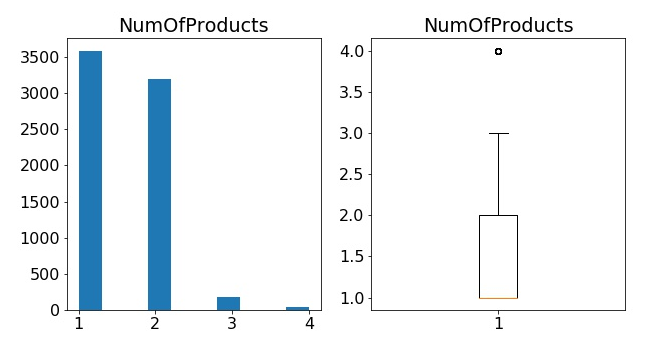
### Es Miembro Activo:

El 51% de clientes es miembro activo.

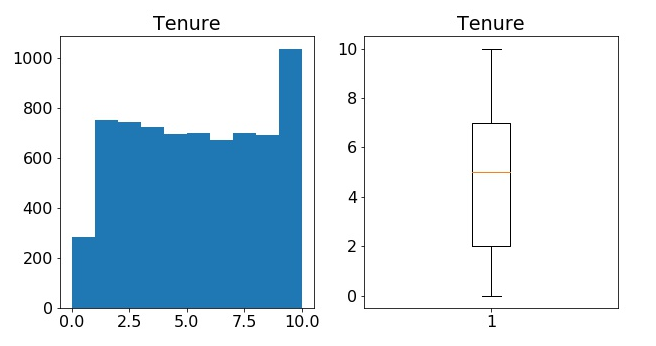


### Número de Productos:

La mayoría de clientes tienen entre 1 y 2 productos, esta variable tiene valores atípicos.



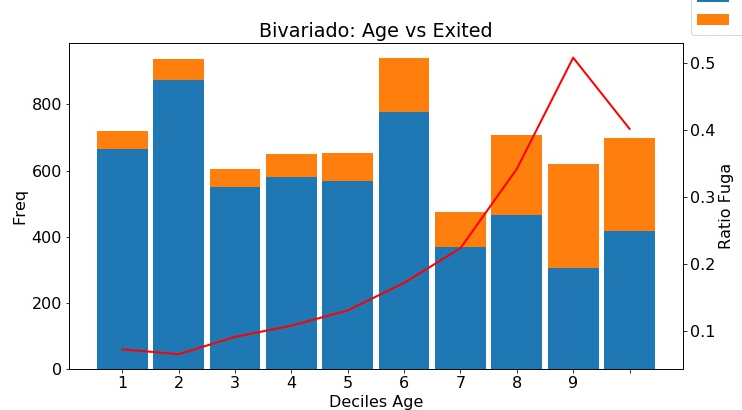
### Permanencia:



# Análisis Bivariado:

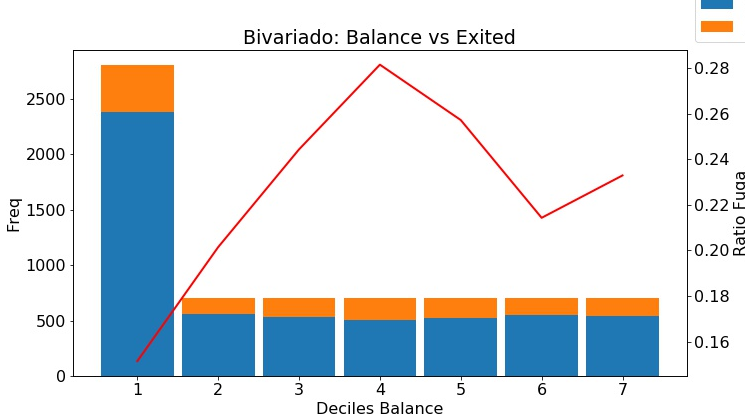
## Año vs Bajas:

Mientras más edad tiene un cliente la probabilidad de fuga es mayor.



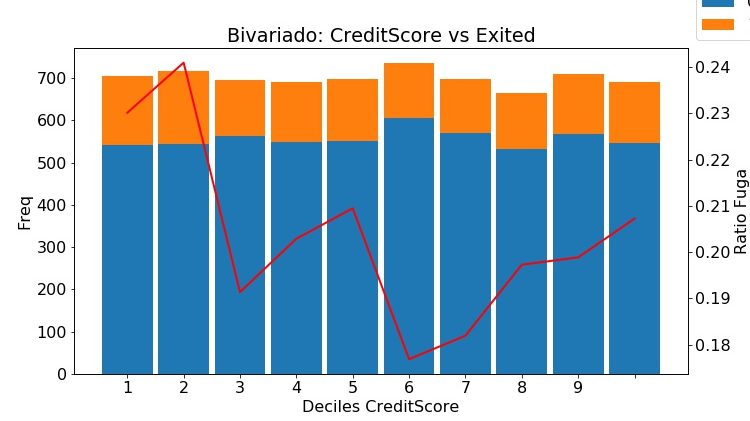
## Saldo vs Bajas:

Mientras más Saldo tiene un cliente, la probalidad de fuga disminuye.



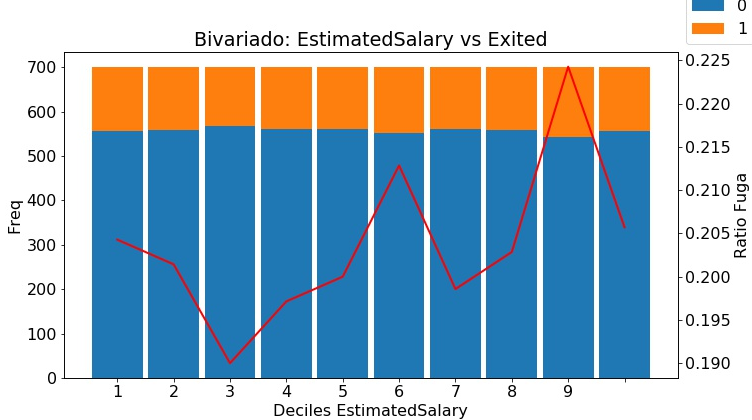
## Score Crediticio vs Bajas:

Ante un mejor score crediticio, menor es la probalidad de fuga, pero no son variables fuertemente correlacionadas, no es una relación esperada.



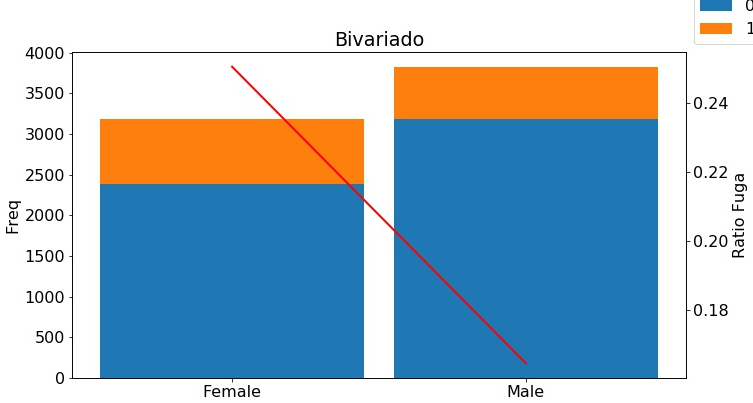
## Salario Estimado vs Bajas:

Son dos variables bajamente correlacionadas.



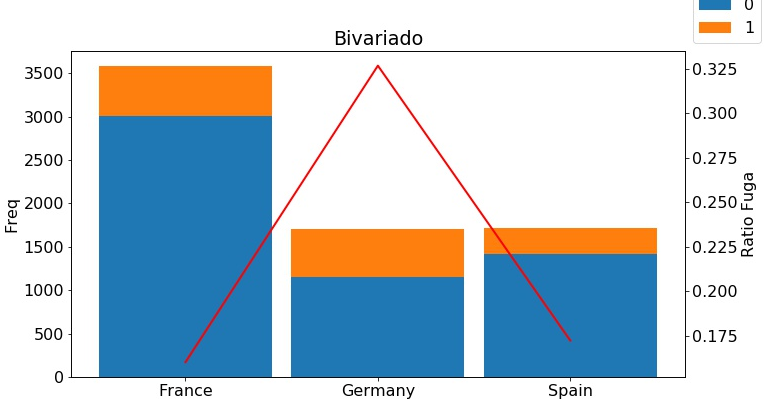
## Bajas por Género:

La diferenciación por generó no aporta a la estimación de la fuga.



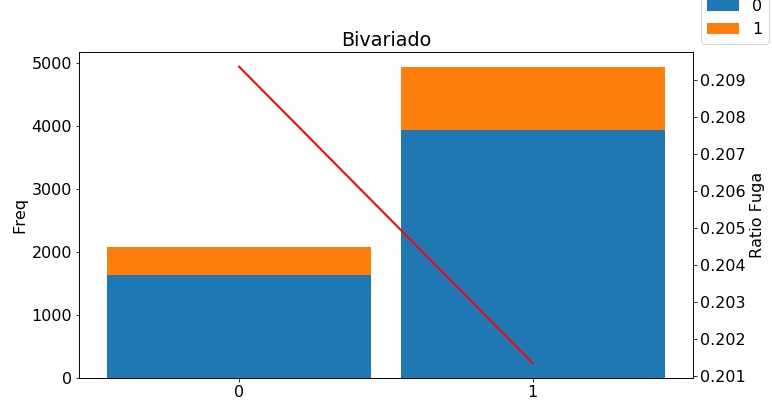
## Bajas por Ubicación Geográfica:

Esta variable si aporta, se puede hacer una segmentación por geografía, porque los índices de fuga si son marcadamente diferentes.



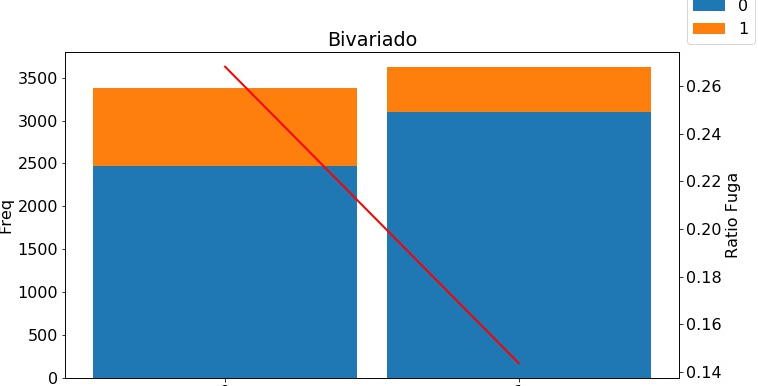
## Bajas por ratio de fuga (tiene crédito):

Esta relación, no aporta al modelado.



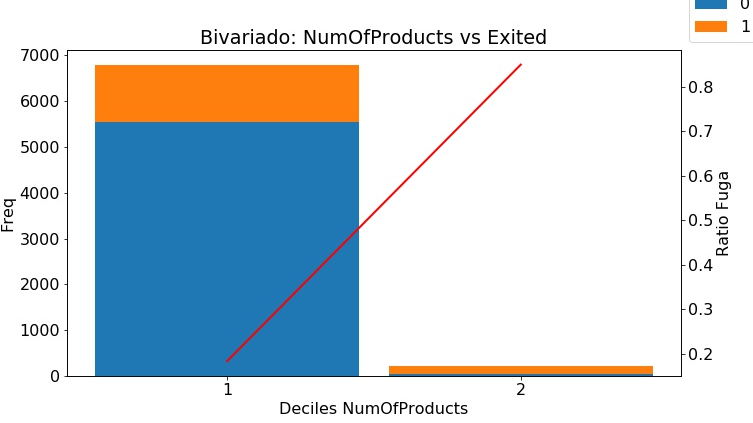
## Bajas por ratio de fuga (es miembro activo):

Esta relación, no aporta al modelado.



## Bajas por número de productos

Esta variable si aporta, se puede hacer una segmentación por cantidad de productos, porque los índices de fuga si son marcadamente diferentes.

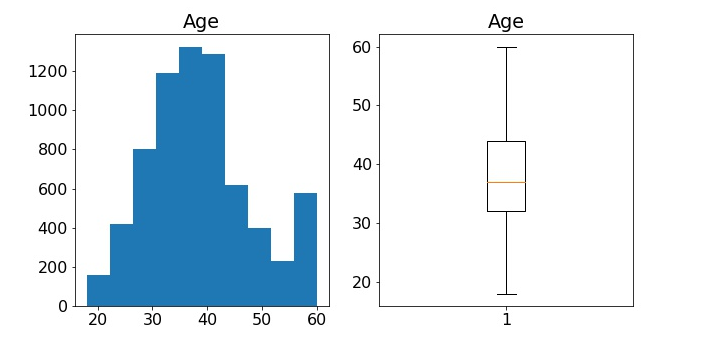


# Tratamiento de Datos:

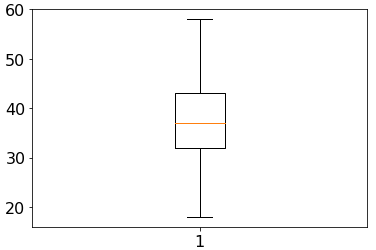
## Edad:

Se verifica que la variable edad tiene valores atípicos de más de 64 años, para lo cual se asignara la edad de 64 años para estos casos:

*ds1.loc[ds1["Age"]>=64,"Age"]=64*



Resultando una variable limpia.

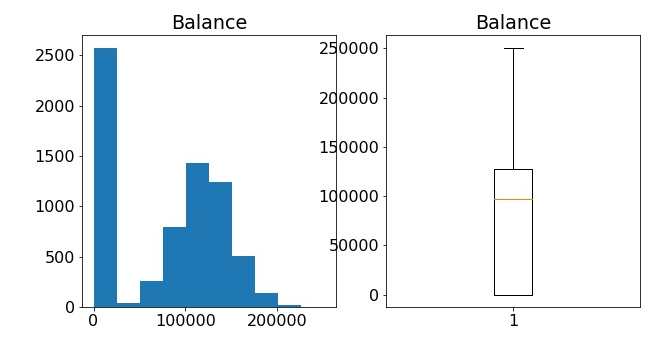


## Saldo:

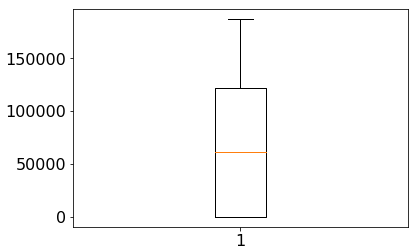
La variable saldo, se tiene que ajustarse, porque hay valores extremos, como el cero y mayores a 185K.

*ds1.loc[ds1["Balance"]>=185160.94680000003,"Balance"]=185160.94680000003*

*ds1.loc[ds1["Balance"]<=53102.854400000004,"Balance"]=53102.854400000004*



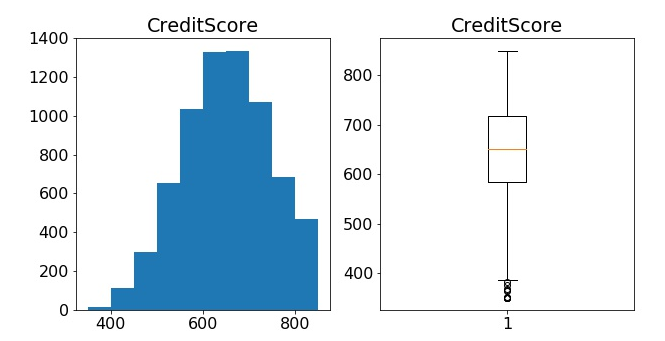
Realizando el tratamiento, esta variable ya esta correcta.



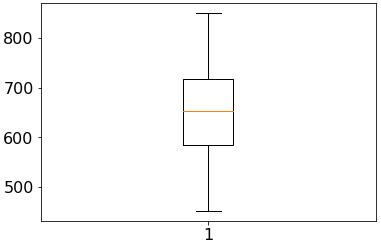
## Score Crediticio:

Esta variable tiene valores atípicos extremos, los cuales son necesarios ajustar.

*ds1.loc[ds1["CreditScore"]<=453,"CreditScore"]=453*



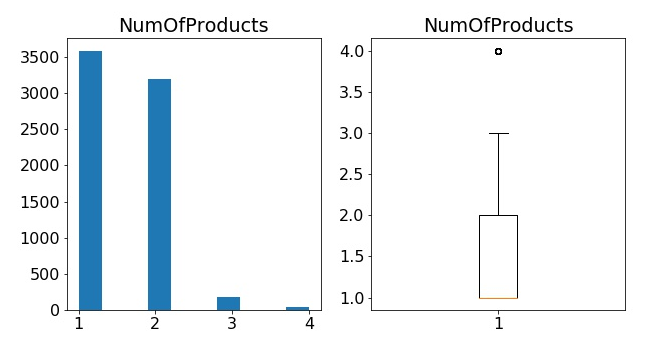
Con el tratamiento de los valores atípicos, la variable esta conforme para entrar el modelo.



## Número de Productos:

Esta variable tiene valores perdidos, por lo cual se tendrá que completar y volverla una variable categórica.

*ds1.loc[ds1["NumOfProducts"]>1,"NumOfProducts"]=0 ds1.NumOfProducts=ds1.NumOfProducts.astype('category')*



# Modelamiento (Fundamentar la(s) técnica escogida, Explicar los pasos realizados , mostrar y comprar la evaluación de resultados).

El análisis a continuación se dividio en dos partes, la primera se trata sobre técnicas de aprendizaje no supervisado, específicamente sobre técnicas de agrupamiento de variables (reducción de dimentiones) y/o registros (análisis clúster). En la segunda parte, se aplicó técnicas supervisadas, que nos permite clasificar a los clientes que abandonarán la empresa y los clientes que se quedarán.

## Análisis no supervisado

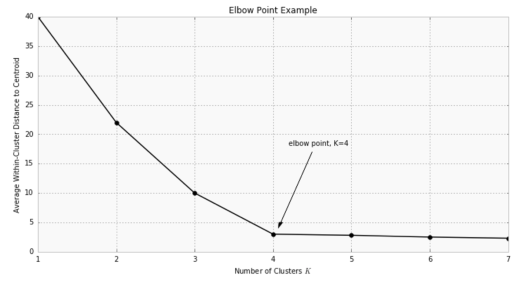
Las primeras opciones que se desarrollaron fueron las técnicas de reducción de dimensiones, como ejemplo Análisis de Componentes Principales, Análisis Factorial, Análisis de Correspondencia, entre otros. Luego, se analizó el agrupamiento natural de los clientes basándonos en las 6 variables registradas a través del análisis Clúster.

### Análisis de Componentes Principales

En el desarrolló se aplicó Análisis de Componentes Principales para determinar las 2 primeras componentes porque concentran el 99% de la varianza explicada, esta reducción nos servirá para poder visualizar en dos dimensiones el análisis clúster.

### Análisis Clúster

Luego, se realizó un análisis clúster de los clientes, el cual nos permitió identificar agrupaciones naturales de acuerdo con las distancias que se generan a partir de las 6 dimensiones. Para determinar el número de clúster se utilizó el diagrama de codo el cual muestra la distancia media entre los puntos de datos y su centroide. Como el valor de la media disminuirá a medida de aumentemos el valor de K, deberemos utilizar la distancia media al centroide en función de K y encontrar el “punto codo”, donde la tasa de descenso se “afila”. Para nuestro ejemplo se obtuvo 4 clúster.



Utilizando las dos dimensiones se visualiza la distribución de los 4 grupo, como se ve a continuación.

Imagen que contiene captura de pantalla

Descripción generada con confianza muy alta

Al realizar el perfilamiento de estos 4 grupos se identificó que en cada clúster se agruparon aproximadamente 1300 clientes. Asimismo, se identificaron variables que variaban significativamente entre clúster, por ejemplo, el salario estimado.

Imagen que contiene captura de pantalla

Descripción generada con confianza alta

## Análisis supervisado

En esta sección se presenta los resultados obtenidos al aplicar diferentes técnicas de aprendizaje supervisado. Estas técnicas pueden dividirse en dos categorías: la regresión y la clasificación. Como el objetivo es clasificar a un cliente que posiblemente nos abandone o no, se aplicarán técnicas de clasificación, entre las principales se aplicaron: regresión logística, árbol de decisión, redes neuronales.

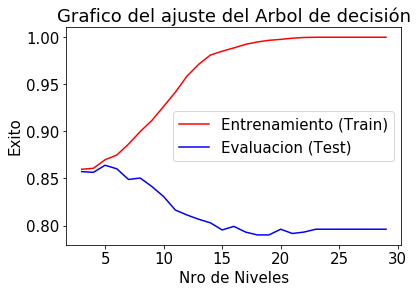
Antes de aplicar los modelos se realizó una separación en datos de prueba y entrenamiento para poder medir los indicadores que evalúan el poder de predicción como son: casos de éxito, fracaso, sensibilidad, especificidad, ROC y GINI.

Previa a la división se realizó una mezcla aleatoria del conjunto de datos. Luego se tomó el 75% de los datos para entrenar y el 25% para evaluar.

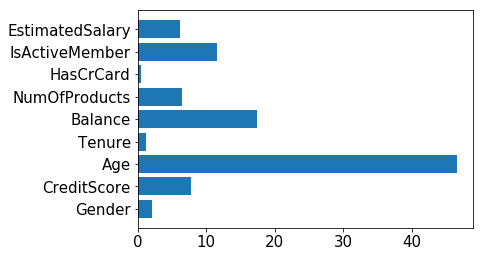
A continuación, se describe los principales hallazgos obtenidos para cada uno de los modelos analizados.

### Árbol de decisión

Luego de realizar múltiples entrenamientos se llega a la conclusión que el nivel más óptimo para el árbol de decisión es 6, ver el gráfico de ajuste del Árbol de decisión.

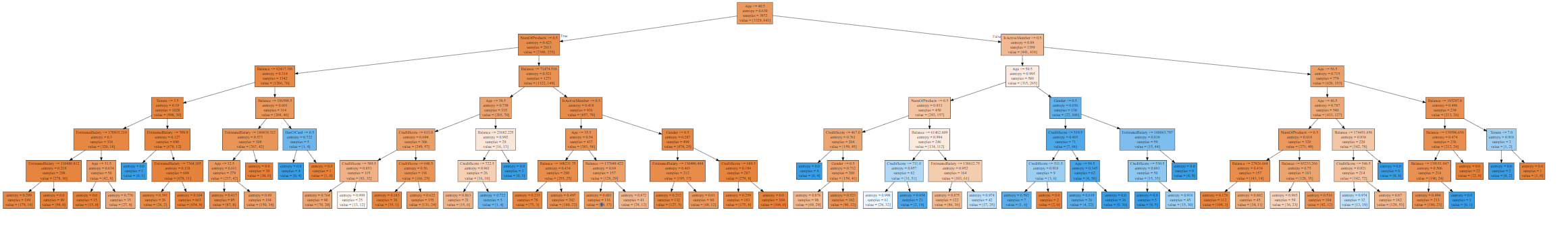


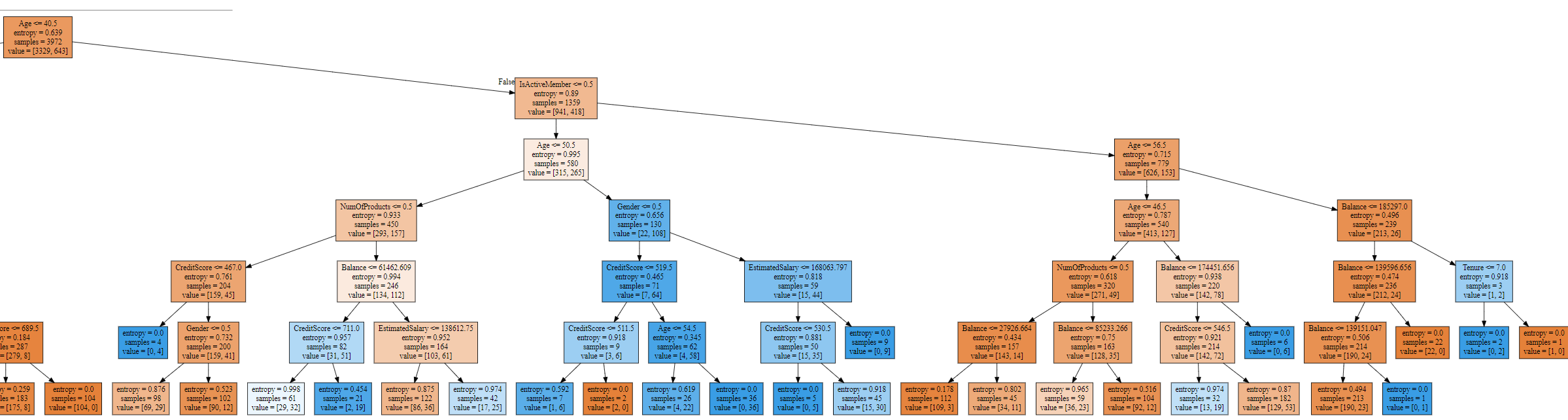
Asimismo, se realizó un análisis de importancia de variables, cuyo resultado se representa en el siguiente gráfico:

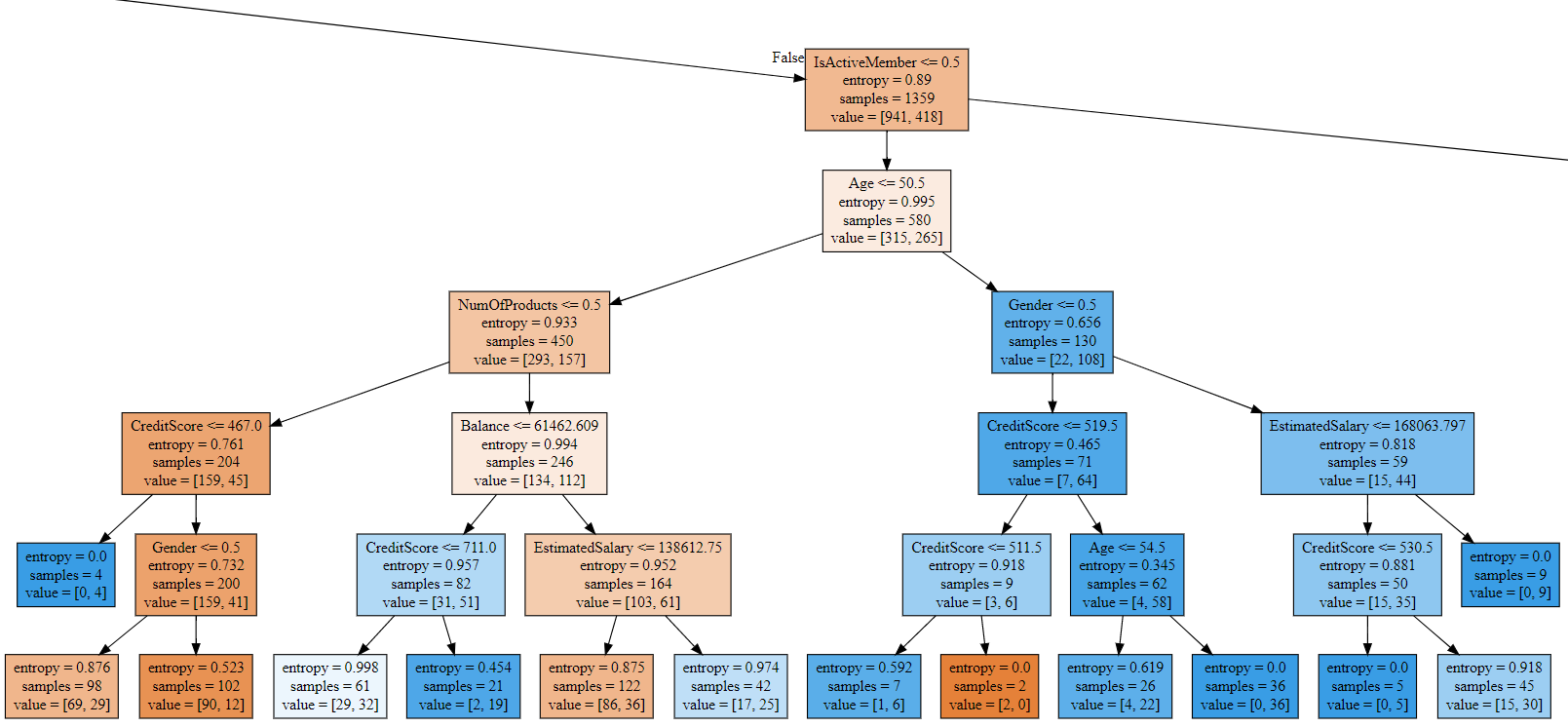


La variable más importante para la clasificación del target es “Age”.

El resultado del árbol de decisión se representa a través del siguiente diagrama el cual presenta los cortes que realizó el árbol a las diferentes variables para que pueda clasificar.







Los principales indicadores que se obtuvo son los siguientes:

Matriz de confusión

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Y predicho | |
| 1 | 0 |
| Y Original | 1 | 1057 | 41 |
| 0 | 144 | 82 |

* Éxito: 86.1%
* Fracaso: 13.9%
* Sensibilidad: 96.3%
* Especificidad: 36.3%
* ROC: 66.3%
* Gini: 32.5%

### Regresión logística

Matriz de confusión

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Y predicho | |
| 1 | 0 |
| Y Original | 1 | 941 | 157 |
| 0 | 150 | 76 |

* Éxito: 76.8%
* Fracaso: 23.2%
* Sensibilidad: 85.7%
* Especificidad: 33.6%
* ROC: 59.6%
* Gini: 19.3%

### Redes neuronales

Matriz de confusión

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Y predicho | |
| 1 | 0 |
| Y Original | 1 | 1077 | 21 |
| 0 | 84 | 142 |

* Éxito: 92.0%
* Fracaso: 8.0%
* Sensibilidad: 98.1%
* Especificidad: 62.8%
* ROC: 75.1%
* Gini: 50.2%

### Combinación de modelos

Es posible mejorar los indicadores del modelo utilizando modelos ensamblados que se pueden clasificar en dos tipos de aproximaciones para la combinación de clasificadores: una aproximación es usar clasificadores similares y combinarlos usando técnicas como bagging, boosting o random forests. Una segunda aproximación es combinar diferentes clasificadores usando model stacking.

Para el presente análisis no se logró implementar, sin embargo, se cuenta con la información necesaria para replicar este tipo de análisis.

### Elección del mejor modelo

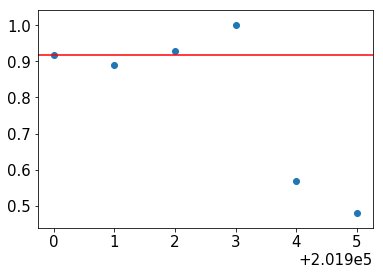
Al comparar los indicadores de éxito, fracaso, sensibilidad, especificidad, ROC y Gini de los 3 modelos desarrollados. A pesar de validar el sobreajuste (overfitting) a través de los múltiples entrenamientos para cada uno de los modelos, es posible aún de generalizar el aprendizaje del modelo final a través de la penalización de los parámetros (Penalización L0, L1, L2, entre otros).

# Seguimiento del modelo

En el último capítulo, se realiza el seguimiento del modelo elegido porque lo más probable es que las condiciones iniciales cambien en el futuro e impacten en el desempeño del modelo. El primer indicador a seguir es la evoluación de la predicción, luego la evolución de la estabilidad de diferentes variables.

## Evolución de la predicción

La predicción parece estable hasta el mes 3, sin embargo, a partir del cuarto mes la predicción cae hasta un 60% y lo mismo ocurre en el 5to mes. Al parecer, se tiene que tomar medidas para identificar el cambio brusco de la tendencia de la predicción. Es posible que se trate de un fenómeno puntual que alteró la predicción al menos por esos 2 meses, si para el 6to y 7mo mes continua con la tendencia negativa se tendría que tomar medidas de reemplazar el modelo, agregando nuevas variables o identificando el porqué del cambio.



## Evolucion de la estabilidad

La evolución de la estabilidad para la variable género, sólo se mantiene en el rango aceptable (0.1 – 0.25) por los dos primeros meses. Estos resultados nos indican que se tiene que revisar e identificar el porqué de la inestabilidad.

### Gender

