Modelo Previsión Ventas Productos Financieros

Memoria

Trabajo Final Master Data Science

Manuel Gonzalez Prados

TFMDS 2020 - 2021

**TABLA DE CONTENIDO**

1. **INTRODUCCION** 
   1. **Motivación**
   2. **Prologo**
   3. **Objetivos**
   4. **Finalidad**
   5. **Interés**
2. **ESTADO DEL ARTE**
   1. **Marco**
   2. **Seguros en España- Cifras**
   3. **Tipos de seguros contratados en España. Evolución anual**
   4. **Tipos de servicio del seguro en España**
   5. **Nivel de contratación de seguros por edades**
   6. **Contratación de seguros por ingresos del hogar**
   7. **Niveles de contratación del seguro de hogar en Europa**
   8. **Conclusiones**
3. **CASO DE USO** 
   1. **Planteamiento**
   2. **Importancia del modelo a resolver**
   3. **Datos**
   4. **Variables utilizadas**
   5. **Modelos**
4. **DESARROLLO Y CONSTRUCCION DEL MODELO PREDICTIVO**
   1. **Información del Repositorio**
   2. **Requisitos técnicos**
   3. **Carga de la Base de Datos**
   4. **Información y extracción de la base de datos**
   5. **Objetivo**
   6. **Limpieza y Unión**
   7. **EDA Análisis Exploratorio**
   8. **Importancia de las variables**
   9. **Preprocesado**
   10. **Construcción del modelo**
5. **Conclusiónes**

**……………………………………………………………………………………………………………………………………**

**1 INTRODUCCION**

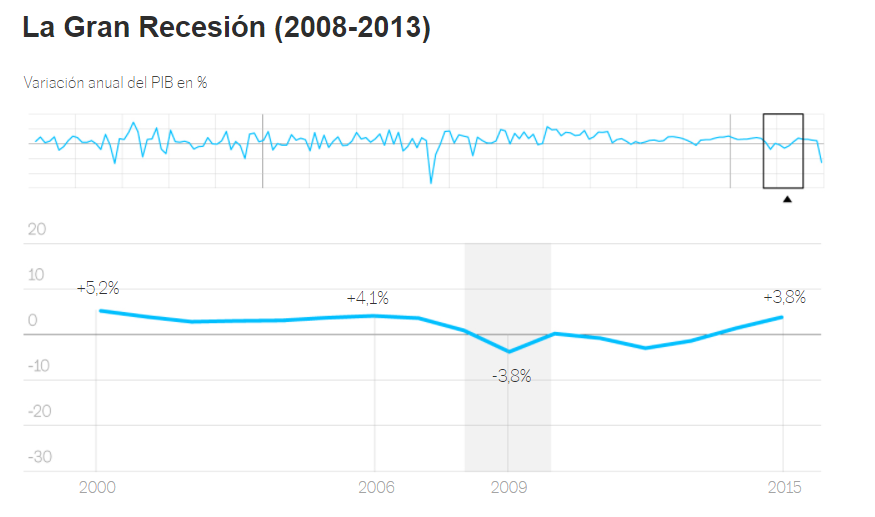
**1.1 MOTIVACION**

Desde 1998 hasta la actualidad, he desarrollado mi vida profesional en una entidad financiera, principalmente en Banca de particulares. Durante estos 23 años he pasado por todas las categorías laborales posibles dentro de una oficina comercial abierta al consumidor. Desde comercial de caja y de mesa, a subdirector y director de oficina. En 2018, motivado por la búsqueda de nuevas habilidades, reciclaje laboral y personal, la adaptación a la nueva realidad de transformación digital y la necesidad de construir un plan alternativo debido a las inciertas perspectivas laborales, huyendo de mi zona de confort decido cursar un Master en Bussines Analytics con la intención de aprender nuevas formas de análisis de negocio y poder ponerlas en práctica. Durante el curso me doy cuenta que aun sin ninguna base de programación o informática, estadística o matemáticas, procediendo de una licenciatura de letras, había encontrado una motivación, una nueva parcela de estudio y un nuevo reto. Decido continuar la formación con el Master en Data Science de K-School, recomendado por un antiguo profesor y siempre avisado de la dificultad técnica del mismo. El resultado lo puedo definir en una frase. ***Intenso pero entusiasmado y con ganas de continuar mi formación.***

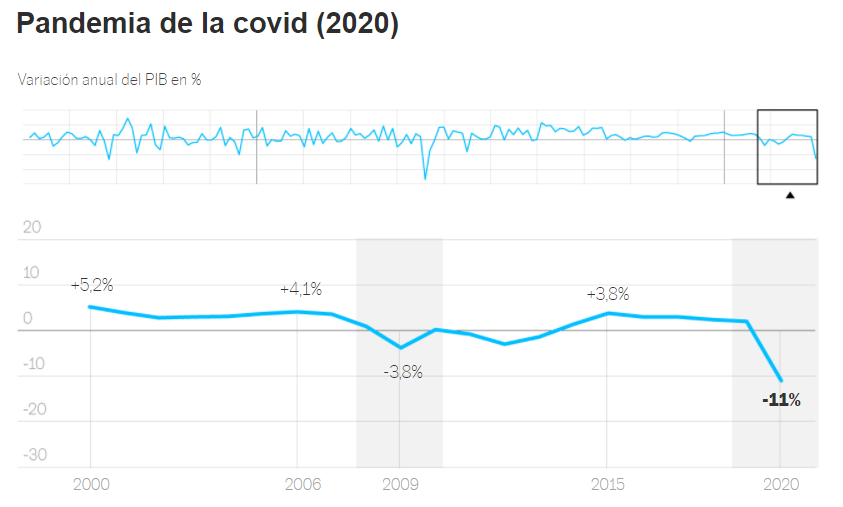
**1.2 PROLOGO**

Los bancos son entidades importantes. Canalizan la riqueza entre los distintos actores de la sociedad. Ese ha sido siempre su principal objetivo y razón de ser. Captar recursos de aquellos que los tienen, remunerarles por esa captación y posteriormente prestarlos recibiendo a cambio un tipo de interés. Se trata de una transacción sencilla y fácil. Las entidades financieras pagan por captar y cobran por prestar. De esta forma todas las partes salen beneficiadas. Quien pone a disposición del banco un dinero que no necesita cobra intereses, el banco cobra intereses al prestarlo y aquellos que reciben el préstamo pueden destinar el dinero para aquello que necesitan. Una empresa, una compra de un vehículo…la compra de una casa… etc. Resumiendo, el negocio tradicional de las entidades financieras ha sido prestar a más tipo de interés el dinero que ha captado.

Hasta 2008 vivimos en un gran periodo de crecimiento económico bruscamente parado por la crisis de las hipotecas subprime. Se inicio un largo periodo de 6 años de crisis profunda y recesión en la que los principales indicadores económicos de la sociedad sufrieron importantes caídas. Las entidades financieras no fueron ajenas a este sufrimiento viendo sus cuentas de resultados mermadas. El negocio tradicional bancario se había paralizado ,se había acabado.



En 2014 se inició la recuperación y cuando parecía que recuperábamos todas las sensaciones, actividad y evidencias económicas de haber olvidado la crisis , en 2020 nos encontramos con una de las peores crisis sanitarias de la historia de nuestra edad moderna generando todavía mayores dificultades y retrocesos que las anteriormente vividas.



Como consecuencia de las diferentes crisis vividas, la reducción de los tipos de interés y de los beneficios obtenidos por el puro negocio tradicional de captar y prestar dinero, este modelo de negocio había dejado de ser el motor principal de la cuenta de resultados de una entidad. La venta de otros productos financieros como los Fondos de Inversión, Planes de Pensión y los seguros de riesgo entre otros, consiguieron acaparar toda la importancia. Hoy en día, la generación de comisiones adheridas a la comercialización de estos productos ha supuesto un vuelco en la estrategia comercial, especializándose y poniendo el foco en su venta mediante el asesoramiento especializado a los clientes por parte de los empleados de las sucursales.

**1.3 OBJETIVOS**

Los Seguros de Riesgo comercializados en las oficinas bancarias, así como su mantenimiento en cartera durante 5 años de media, son de gran importancia dentro de la cuenta de resultados de una oficina y por extensión de un banco. En este escenario y a través de un conjunto de datos pertenecientes a 450.000 clientes he querido desarrollar un **modelo predictivo de compra de estos productos financieros concretando en los Seguros del Hogar**.

**1.4 FINALIDAD**

Generar un modelo predictivo de clasificación que ayude a toda la fuerza comercial de las sucursales a orientar la comercialización, a optimizar los tiempos, metodologías y sistemas utilizados. Todo ello en busca de un mayor éxito de ventas, generación de margen económico y satisfacción de los clientes.

**1.5 INTERES**

El interés de esta investigación se centra en tres niveles; Empresarial, Comercial y Optimización Comercial.

* Intereses Empresariales: El principal objetivo de una entidad financiera, como cualquier otra empresa privada es maximizar el beneficio que sus socios han invertido, y por los cuales, esperan un retorno a través del reparto de dividendos. La generación de comisiones por la venta de seguros es una de las vías más importantes para generar beneficios y en consecuencia una mayor cuenta de resultados.
* Desde el punto de vista comercial, la competencia a nivel de mediación y venta de seguros en grandísima. No solo las propias entidades aseguradoras, sino que todas las entidades financieras de hoy en día venden seguros. El hecho de asesorar, vender y captar a un nuevo cliente a medio largo plazo es un hito importante que hay que potenciar.
* Por último, hablar de la optimización del trabajo. No solo hay que trabajar duro, sino trabajar de forma eficaz y eficiente. La jornada laboral de un empleado de banca se resume en Asesorar y vender. Hay que vender y hay que llegar a cuantos más clientes mejor. Si todo esto lo hacemos centrando la llamada, optimizando el tiempo, sabiendo a quien llamamos, a quien nos dirigimos y porqué, mejoraremos en todos los sentidos esa eficiencia y eficacia buscada.

**2. ESTADO DEL ARTE**

**2.1 MARCO**

Las Entidades Financieras y los seguros, son el principal sustento de nuestro sistema financiero y asegurador. La economía de cualquier país tiene como pilares principales a estos dos sectores. Tanto la Banca como las aseguradoras están en constante desarrollo tecnológico y organizativos. Ambos sectores jugarán siempre un papel principal en cualquier crisis presente o futura ya que en ellos están depositados planteamientos de recuperación solvencia y liquidez para minimizar el impacto económico en empresas y hogares.

Algunos datos de 2020. En los primeros meses lógicamente influenciada por la gran crisis sanitaria que vivimos, se redujo la siniestralidad en ramos como autos o viajes y aumentó en ramos de decesos, impago de alquileres, vida y salud.

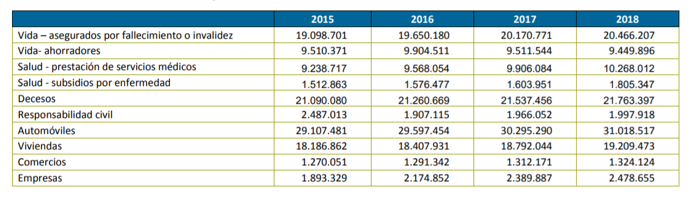
A nivel empresarial se ha envidenciado una alta tasa de anulación de polizas. Pymes y autónomos son sectores que han padecido con mayor intensidad esta pandemia teniendo como consecuencia cierre de comercios y empresas.

Motivado por todo ete descenso asegurador, las primas has subido en las renovaciones y se han incrementado riesgos que antes no había. Ej. riesgos cibernéticos y medioambientales.

Predicciones para seguros en 2021. La **escalada de precios en el mercado global de seguros** se hará evidente. Daños Materiales, Responsabilidad Civil, D&O y Ciberriesgos serán las líneas de negocio más afectadas. **La carga de siniestros relacionados con la pandemia se reducirá** progresivamente debido a la exclusión de las coberturas pandémicas en los acuerdos objeto de renovación

**2.2 SEGUROS EN ESPAÑA – CIFRAS**

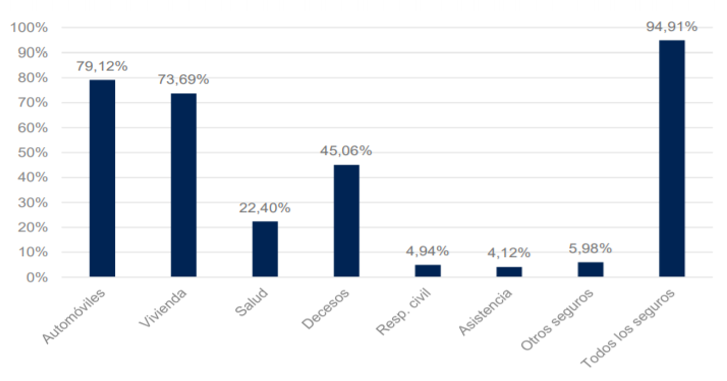
Todos los años el INE publica un detalle de los gastos y presupuestos de las Familias españolas. Entre ellos están lo que se gasta en seguros detallándolo por familia exceptuando los de vida que no se informan.

**2.3 TIPOS DE SEGUROS CONTRATADOS EN ESPAÑA. EVOLUCION ANUAL**

**Nivel de contratación de seguros por parte de los hogares**.

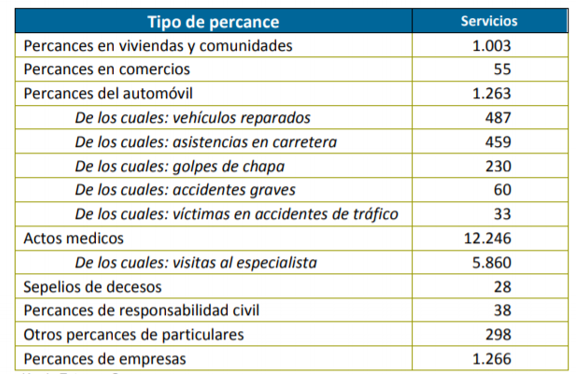
Actualmente estamos censados en España 46.722.980 habitantes, contando con un parque de seguros que supera los 119.000.000 de seguros a 2018. El número de seguros de Hogar- Vivienda asciende a más de 19.000.000. Un 15% del total.

El aseguramiento de las viviendas españolas tiene un crecimiento continuo. Mas de 1MM de seguros de hogar se han contratado en el periodo de 2015 a 2018. Seguros de Auto y de Vivienda son los que más crecimiento han tenido.



Fuente: <https://www.unespa.es/que-hacemos/publicaciones/>

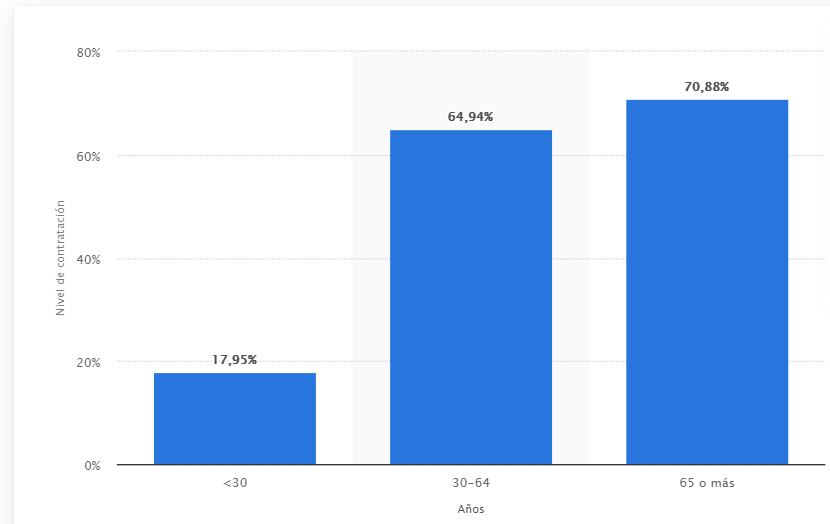
**2.4 TIPOS DE SERVICIO DEL SEGURO ESPAÑOL**



Fuente: <https://www.unespa.es/que-hacemos/publicaciones/>

Quitando los servicios médicos, supone el 10% de la demanda de servicios aseguradores.

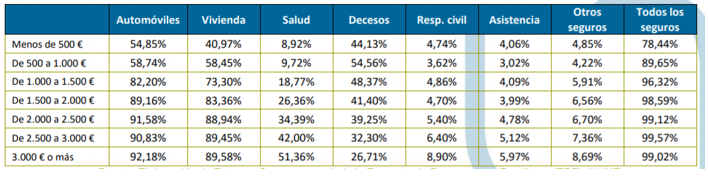
**2.5 NIVEL DE CONTRATACION DE SEGUROS POR EDADES**



Fuente: <https://es.statista.com/estadisticas/967761/nivel-de-contratacion-de-seguros-de-hogar-por-edad-espana/>

Esta grafica visualiza como el seguro de hogar es contratado a partir de los 30 años. Claramente influenciado por el modelo de vida y dificultad de acceso a la vivienda de las personas jóvenes. Esto lo podremos ver en nuestro modelo predictivo.

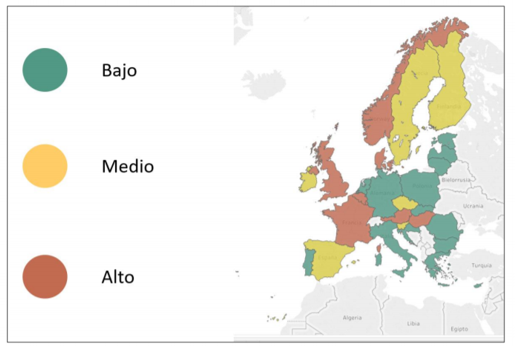
**2.6 CONTRATACION DE LOS DISTINTOS SEGUROS POR INGRESOS DEL HOGAR**



Las contrataciones de los seguros de hogar dependen en gran medida de los ingresos familiares. Se evidencia un gran corte en los hogares. Las contrataciones despegan con los hogares donde se ingresan mínimo 1000€.

Por último, visualizar a nivel Europeo las diferencias de gasto en los seguros de hogar.

**2.7 NIVELES DE CONTRATACION SEGURO HOGAR EN EUROPA**



**2.8 CONCLUSIONES**

El seguro de hogar está muy arraigado en la sociedad española. La cantidad del presupuesto del hogar destinada a los seguros como en el caso de cualquier otro componente de dicho gasto, depende de muchos factores estructurales y contextuales de cada hogar, como el lugar de residencia, la situación laboral de la familia, la escala salarial de sus integrantes, etc.

**Modelo.** Con los datos y variables disponibles intentaremos entrenar una solución comercial que facilite la venta de este producto a los asesores de las entidades financieras.

Dentro del sector financiero existen múltiples modelos de clasificaciónrelacionados con las distintas necesidades.

* Modelos de Riesgo
* Modelos de Morosidad
* Modelos de RRHH
* Modelos de Segmentación Clientes
* Modelos de ventas FFII – PP – Seguros – Servicios Financieros
* Modelos de clientes digitales

**3 CASO DE USO**

* 1. **Planteamiento de Caso de Uso.**

Modelo predictivo contratación de seguros de hogar

**3.2 Importancia del problema a resolver**

La generación de comisiones es fundamental para la cuenta de resultados de la oficina y del banco. Los seguros de hogar se quedan en cartera durante un periodo medio de 5 años. Cada seguro de hogar contratado deja una comisión directa del 15 %. Esto sobre un seguro de hogar de prima media de 300€ supone 45€ de comisión anual por cliente. Nuestra base de datos correspondiente únicamente a 162 oficinas y 450.000 clientes podría llegar a generar unas comisiones anuales de más de 17mm€. Solo con esta cifra y extrapolándola a un colectivo de 3 – 4 millones de clientes, queda más que explicado la evidente y clara la necesidad de identificar potenciales clientes que sean susceptibles de contratar el seguro de hogar.

* 1. **Datos utilizados.**

Los datos provienen de la suma de diferentes conjuntos de información obtenidos directamente de la entidad financiera. TODOS LOS DATOS HAN SIDO ANONIMIZADOS. SE HAN ELIMINADO LOS NOMBRES Y EL NÚMERO DE CLIENTES INTERNOS, SE HAN ELIMINADO EL NÚMERO Y LAS ESPECIFICACIONES DE LAS DIRECCIONES DE ÁREA Y DE OFICINA Y, POR ÚLTIMO, SE HAN ELIMINADO LOS NOMBRES DE LOS ASESORES COMERCIALES. A todos estos datos se les ha asignado un número de identificación ficticio y secuencial, quedando únicamente los datos de tenencia o no de producto por parte de los clientes.

* 1. **Variables utilizadas**
* ***DZ***. Identificación Dirección de Zona a la que pertenece la oficina. Una Dirección de Zona engloba varias oficinas. Total 11 Direcciones de Zona
* ***OFICINA***. Numero de Oficina / Sucursal de banco
* ***CLIENTE***. Numero de cliente
* ***EDAD***. Edad del cliente
* ***ESTA***\_***CARTERIZADO***. Identifica si el cliente pertenece o no a una cartera.
* ***CARTERA\_PATRON.***  Tipo de cartera a la que pertenece el cliente.
  + Asesoramiento Financiero
  + Tutela. Familiar de cliente Asesoramiento Financiero
* ***CLIENTE\_BBP.*** Cliente con saldos superiores a 500.000€ identificado como colectivo Banca Privada.
* ***GESTOR***. Numero identificación del gestor/ Asesor Financiero de la sucursal
* ***TIP\_GESTOR***. Tipo de gestor. Figura de una oficina que ofrece el asesoramiento.
* ***CODIGO\_CARTERA***. Numero identificación cartera a la que pertenece el cliente.
* ***MARCA\_AF\_CCTE***. Identifica si el tipo de gestor es de oficina u Online.
  + AF- Asesor Financiero (Oficina)
  + CCTE – Gestor Online
* ***MARCA\_BANCA\_PERSONAL***. Cliente perteneciente a cartera Asesoramiento Financiero e identificado como colectivo Banca Personal.
* ***SEGMENTO\_RECORRIDO***. Identifica el potencial recorrido comercial del cliente para una mayor vinculación.
  + Alto Recorrido
  + Medio Recorrido
  + Bajo Recorrido
* ***SEGMENTO\_VALOR.*** Valor del cliente
  + Alto ( Alto nivel Patrimonial o alto nivel de vinculación )
  + Medio ( Clientes que sin cumplir el anterior requisito tienen un alto nivel de fidelización.
  + Bajo ( Resto de clientes )
* ***CAMINO\_DIGITAL***. Se diferencian 4 tipo de clientes según la utilización de canales digitales.
  + Comprador
  + Consultivo
  + Transaccional
  + Poco uso
* ***DIGITAL\_3\_MESES***. Identifica si el cliente ha utilizado medios digitales durante los últimos 3 meses.
* ***LP\_DOMIC\_INGRESOS.*** Tiene o no tiene ingresos domiciliados
* ***LP\_OFIC\_INTERNET.*** Tiene o no tiene servicio internet
* ***LP\_REC\_LTGA\_OTR.*** Tiene o no tiene recibos domiciliados
* ***LP\_SEG\_ACCIDENT.*** Tiene o no tiene seguro accidentes contratado
* ***LP***\_***SEG\_AUTO.*** Tiene o no tiene seguro automóvil contratado
* ***LP\_SEG\_MEDICOS.*** Tiene o no tiene seguro salud privado contratado
* ***LP\_SEG\_MULTIRRIES.*** Tiene o no tiene seguro hogar contratado
* ***LP\_SEG\_VIDA.*** Tiene o no tiene seguro vida contratado
* ***LP\_TARJ\_CREDITO.*** Tiene o no tiene tarjeta crédito pago fin de mes contratada
* ***LP\_TARJ\_REVOLVING.*** Tiene o no tiene tarjeta crédito pago fraccionado contratada
* ***SF\_AH\_CAPTACION\_TT.*** Saldo en cuenta de ahorro
* ***SF\_FINANCIACION\_TT.*** Importe financiación en activo.
* ***SF\_FONDOS\_INVER.*** Saldo en Fondo de Inversión
* ***SF\_PLAN\_PENSION.*** Saldo Plan de Pensión
  1. **Modelos Machine Learning Utilizados.**
* Modelo Clasificación Regresión Logística
* Modelo Clasificación K-Nearest Neighbor
* Modelo Clasificación TREE
* Modelo Clasificación Xgboost
* Modelo Clasificación Landon Forest - MODELO FINAL ELEGIDO

1. **DESARROLLO Y CONSTRUCCION DEL MODELO PREDICTIVO**

**4.1 Información del Repositorio**

Toda la información de este TFM ha quedado recogida en un repositorio de GitHub al cual se accede a través de la siguiente dirección.

<https://github.com/romayana/Financial-Product-Sales-Forecast-Model.git>

El repositorio se estructura en 7 carpetas y 4 archivos. ( Según posición en repositorio )

Carpeta 1 – Códigos Python Limpieza y Unión

Carpeta 2 – Códigos Python EDA Análisis Exploratorio

Carpeta 3 – Frontend. Aplicación creada para nuestro modelo (APP CallorNot.)

Carpeta 4 – Imágenes .png guardadas de cada una de las gráficas construidas

Carpeta 5 – Memoria. Documento y notebooks memoria.

Carpeta 6 – Códigos Python Modelos clasificación utilizados.

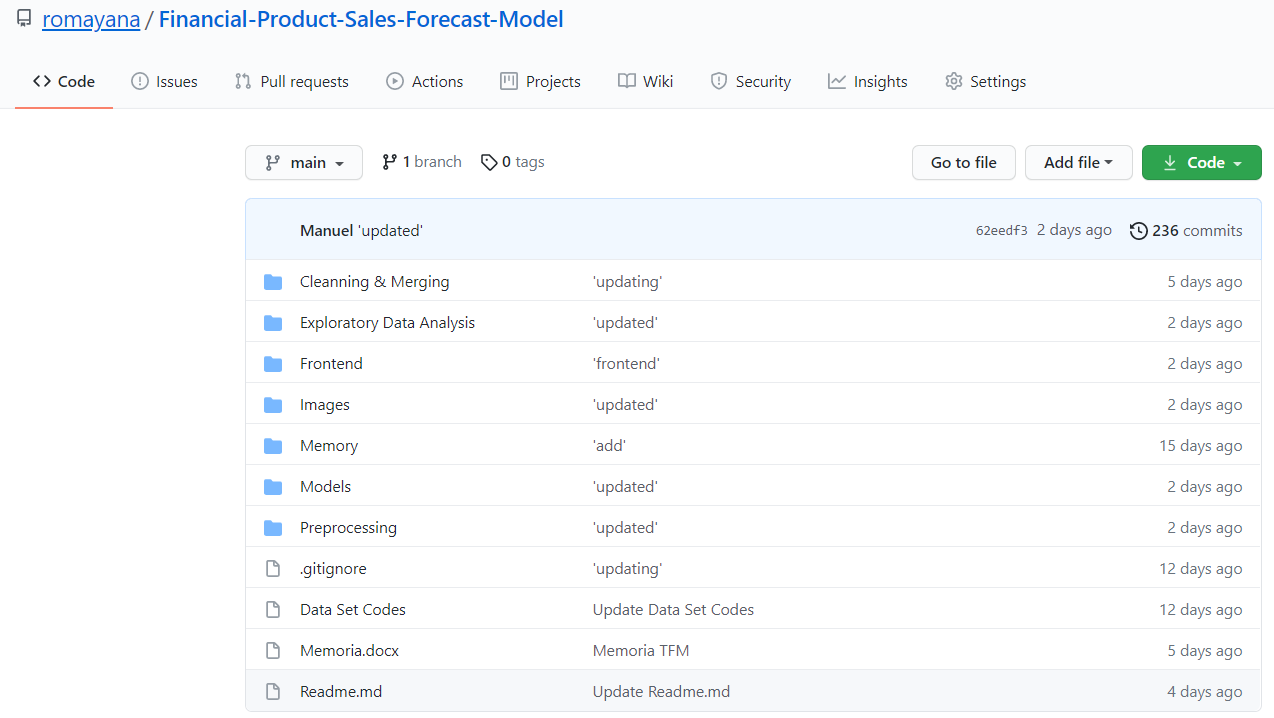
Carpeta 7 – Códigos Python Preprocesado

Archivo 1 – archivo .gitignore. Archivos descartados en las actualizaciones del repositorio

Archivo 2 – Diccionario e información del significado de las variables

Archivo 3 – Memoria TFM.doc

Archivo 4 – Readme con primera información del Trabajo y comunicación de expectativas



**4.2 Requisitos Técnicos**

Para ejecutar los códigos es necesario tener instalado Python versión 3.8 así como distintos paquetes o librerías. Se recomienda tener instalada la Suite Anaconda donde se encontrarán preinstalados la mayoría de los paquetes y librerías que son necesarias.

Librerías utilizadas. La mayoría ya precargadas en Suite Anaconda

# Librerías manejo y análisis de estructuras de datos.

import pandas as pd

# Librerías especializada en el cálculo numérico y el análisis de datos

import numpy as np

# Librerías de Métricas

from sklearn.metrics import f1\_score, recall\_score, precision\_score, accuracy\_score

from sklearn.metrics import roc\_auc\_score, roc\_curve

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import classification\_report

from sklearn.metrics import auc

from sklearn.model\_selection import KFold

from sklearn.metrics import log\_loss

from sklearn.model\_selection import cross\_val\_score

# Librerías de Visualización

import matplotlib.pyplot as plt

import pylab as pl

import seaborn as sns

from pylab import rcParams

from matplotlib import pyplot

# Librerías de Modelos

from sklearn.linear\_model import LogisticRegression

from sklearn.neighbors import KNeighborsClassifier

from sklearn.neighbors import KNeighborsRegressor

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.model\_selection import train\_test\_split

from collections import Counter

from imblearn.over\_sampling import SMOTE

from imblearn.under\_sampling import NearMiss

Adicionalmente será necesaria la instalación de las siguientes librerías.

* Imbalanced learn – proporciona herramientas cuando se trata de la clasificación con clases desequilibradas. (Instalación mediante consola - pip install imbalanced learn )
* pydotplus - Visualización de árboles de decisión en Python con PyDotPlus. ( instalación mediante consola - pip install pydotplus )
* streamlit – creación e intercambio de aplicaciones web personalizadas para el aprendizaje automático y la ciencia de datos. ( instalación mediante consola - pip install setreamlit )
  1. **Guía de Ejecución y Carga de la base de Datos**

1. Clonar repositorio GitHub <https://github.com/romayana/Financial-Product-Sales-Forecast-Model.git> en carpeta local elegida.
2. Descargar base de datos:

A pesar de haberse **Anonimizado** toda la base de datos, se ha decidido que la misma no estará disponible en el repositorio de GitHub. Para acceder a la base de datos ubicada en el Google Drive del propietario del TFM, se tendrá que solicitar permiso y acceso a la misma dirigiendo correo electrónico a [manuelgonzalezprados@gmail.com](mailto:manuelgonzalezprados@gmail.com) el cual previa valoración de los fines y objetivos perseguidos podrá compartir el enlace con la persona solicitante.

Una vez compartido el acceso, descargar y ubicar la carpeta entera llamada **Origin\_Data** dentro de la carpeta carpeta local donde se ha clonado el repositorio junto con el resto de carpetas.

1. Ejecutar código con la siguiente secuencia y orden. Los archivos csv se irán guardando en cada una de las carpetas.
   * 1º Carpeta Leasing & Merging
     + \_merging\_data,ipynb
     + \_cleanning\_data.ipynb
   * 2º Carpeta Exploratory Data Analysis
     + EDA.ipynb
   * 3º Carpeta Preprocessing
     + Preprocessing.ipynb
   * 4º Carpeta Models
     + Ejecutar los modelos.
   * 5º Carpeta Frontend
     + Aplicación Callornot.

Dentro de la carpeta Memory existe la posibilidad de ejecutar el código a través del notebook memoria & código. El modelo final al que se hace referencia es Random Forest.

* 1. **Información y extracción de la base de datos**

La información y explicación detallada de las variables que conforman la base de datos se encuentra dentro de una de las carpetas de este repositorio con el nombre de Data-Set-Codes. También ha quedado explicada dentro de esta memoria en la sección **3.4 Variables utilizadas.**

La extracción de los datos se ha realizado desde un sistema de información de gestión de una entidad financiera. Sistema de información que guarda millones de datos de tipo financiero y económicos.

Toda la información y datos necesarios para el estudio del modelo de clasificación se han obtenido de forma directa y con permisos limitados de una entidad financiera real. Permisos limitados ya que no se ha podido disponer de mucha información que hubiese mejorado el modelo. Información como por ejemplo género, estado civil, hijos, clase económica, renta disponible, importe de nómina, detalle de compras realizadas, detalle de llamadas comerciales realizadas y otras muchas variables.

Para la construcción de la base de datos final se han ido descargando de este sistema de información de gestión y de forma manual, archivos individuales extensión xlsx, relacionados con distintos epígrafes como saldos en cuenta, saldos en fondos de inversión o planes de pensión, líneas de producto, tarjetas, seguros o tipo de segmentación. En total han sido 99 archivos Excel descargados. 9 archivos por cada una de las 11 Direcciones de Zona disponibles lo que ha generado finalmente una base de datos de 15 millones de datos de 450.000 clientes.

El peso total de los datos originales es de 112 MB (117.706.752 bytes) los cuales después de haber sido limpiados se han quedado en 70,3 MB (73.742.213 bytes).

**La carpeta origin\_data está compuesta por las distintas carpetas y archivos Excel descargados ya anonimizados. Los datos se han anonimizado** previamente al guardado en esta carpeta. Numero de identificador de cliente ha sido cambiado por una secuencia desde 1 a 450.000. Números y códigos distintivos de las Direcciones de Zona han sido cambiados por una secuencia del 1 al 11. Nombres , direcciones y números de identificación fiscal de los clientes han sido eliminados.

Detalle Carpeta con Datos Originales. 11 carpetas x 9 archivos.



* 1. **Objetivos**

A través de un conjunto de datos pertenecientes a 450.000 clientes he querido desarrollar un modelo predictivo de potencial compra de estos productos financieros, concretando en el Seguro de Hogar. Un modelo predictivo de clasificación que ayude a toda la fuerza comercial de las sucursales a orientar la comercialización, a optimizar los tiempos, metodologías y sistemáticas utilizadas. Todo ello en búsqueda de un mayor éxito de ventas satisfacción de los clientes y generación de margen para la entidad financiera.

Mi objetivo final será implementar una aplicación ( CallorNot) donde mediante un cuestionario, incorporando una serie de características de un cliente podamos predecir la posibilidad de que ese cliente sea susceptible de contratar o no un seguro de hogar, sugiriendo finalmente si llamar o no llamar al cliente.”

* 1. **Unión de archivos y Preparación de los Datos**
* Unión. 11 carpetas correspondientes a 11 Direcciones de Zona y 9 archivos Excel cada uno se fusionan en un solo Data Frame.

De forma secuencial se han ido leyendo los archivos de cada una de las Direcciones de Zona creando una única lista agregada por DZ y finalmente uniendo en una sola base de datos la totalidad de los 99 archivos Excel originales individuales.

Todo el proceso de unión está ubicado en el notebook \_merging\_data.ipynb al cual se puede acceder dentro de la carpeta Cleanning & Merging del repositorio.

* Preprocesado y Limpieza de los datos. Esta tarea de preprocesado de los datos o sencillamente de preparación de los datos, la iremos realizando a lo largo de nuestro estudio y en distintas secciones . Detectaremos Nans o valores nulos, se corregirán, se buscarán posibles outliers en variables, se buscarán las mejores o más importantes variables mediante técnicas de feature selection, se estudiará la posible reducción de dimensionalidad, convertiremos variables categóricas en numéricas y normalizaremos las variables a una escala común. Todo ello con el fin de construcción un data set de calidad y así poder trabajar con la mejor información de datos para la construcción definitiva del modelo.

Inicialmente utilizamos la base de datos resultado del anterior proceso de unión. Esta base de datos cuenta con 451.374 filas y 30 columnas.

La primera decisión que tomamos con el data set recién construido y unido es estudiar y visualizar posibles deficiencias con la detección de valores Nans. De 30 variables, encontramos 10 que contienen Nans. Dado el conocimiento que tenemos de la base de datos, estas se rellenan con distintos valores acordes a la categoría y segmento del dato.

Este primer proceso de limpieza está ubicado en el notebook \_cleanning\_data.ipynb al que se puede acceder dentro de la carpeta Cleanning & Merging de este repositorio.

* 1. **EDA. Análisis Exploratorio.**

Comenzamos el análisis de nuestra base de datos analizando primero el Marco de Datos. Nuestro objetivo es realizar un análisis exploratorio, estudiando los datos, buscando posibles patrones, visualizando los datos estadísticos y encontrando posibles relaciones que serán útiles para entender el contexto del Marco de Datos y posteriormente para nuestro modelo de clasificación.

Esperamos obtener información básica del marco de datos y una información más profunda sobre nuestro Objetivo Seguro de Hogar

Disponemos de una base de datos consistente en 451.374 filas, 30 columnas - variables y 13.541.220 datos. Total memoria utilizada 103.3MB. El nombre de las variables son las siguientes:

'cliente', 'saldo captación', 'esta\_carterizado', 'cliente\_bbp', ‘tipo\_gestor', 'gestor', 'cartera\_patron', 'codigo\_cartera', 'digital\_3\_meses', 'camino\_digital', 'saldo\_financiacion', 'saldo\_ffii', 'saldo\_plp', 'lp\_dom\_ingresos', 'lp\_tjta\_cto', 'lp\_tjt\_rev', 'lp\_rbos', 'lp\_of\_int', 'marca\_bp', 'marca\_ccte', 'edad', 'seg\_valor', 'seg\_recorrido', 'dz', 'oficina', 'lp\_seg\_vida', 'lp\_seg\_acc', 'lp\_seg\_salud', 'lp\_seg\_hogar', 'lp\_seg\_auto'

Estas variables las hemos juntado por categorías y segmentado en 6 bloques distintos para poder estudiarlas mejor.

* 1. Bloque Unidades de Negocio
  2. Bloque Edad de los clientes
  3. Bloque Ahorro y Financiación
  4. Bloque Servicios
  5. Bloque Seguros de Riesgo
  6. Bloque Segmentación

#  ***Contexto.***

Para poder entender la importancia de nuestro modelo, el impacto económico y repercusión que puede llegar a tener el hecho de saber diferenciar a los clientes susceptibles de contratar el seguro de hogar vamos a explicar haciendo un retrato piramidal , cómo está estructurada la entidad financiera.

* + - 1. Dirección Banca Particulares. Destinada a la atención de clientes particulares / personas físicas.
      2. Direcciones Territoriales. División en un número determinado de Direcciones Territoriales según distribución nacional.
      3. Direcciones de Zona o de negocio. Cada una de las Direcciones Territoriales está dividida en Direcciones de Zona dando cobertura a cada una de las zonas geográficas de ese territorio.
      4. Oficinas. Cada Dirección de Zona está dividida en un número concreto de Oficinas atendiendo a situación geográfica vinculada con esa Dirección de Zona.
      5. Asesores Financieros- Cada una de estas oficinas dispone de un número determinado de Asesores Financieros. Entre 1 y 4. Dependiendo el volumen de clientes.
      6. Clientes de cada una de las Oficinas.

**Nuestra base de datos hace referencia a una de esas territoriales y su estructura interna con un total de 450.000 clientes.**

Teniendo en cuenta este volumen de negocio de clientes, el impacto y ganancia económica de una entidad financiera con 2000 - 3000 oficinas puede ser enorme.

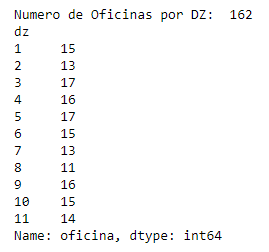
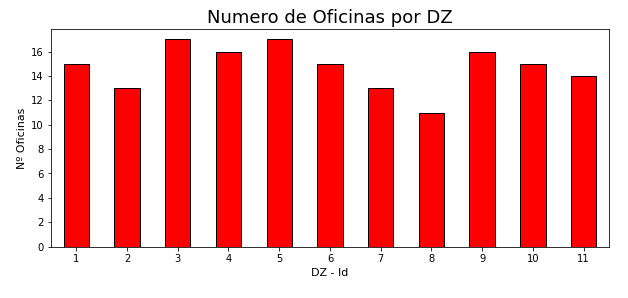
# ***Bloque 1 Variables de Unidad de Negocio.***

Informamos de inicio de cómo queda la estructura de negocio de este retrato piramidal. Se trata de unidades de negocio por lo que no estarán presentes en nuestro modelo de clasificación. Queremos clasificar a los clientes por su propensión al consumo y no por su ubicación territorial.

* Encontramos 11 Direcciones de Zona con 162 oficinas y 451.374 clientes. Estos clientes están asesorados comercialmente por 458 Asesores Financieros. El 90 % de los clientes pertenecen a una cartera de negocio.

En general igualdad de oficinas por DZ.

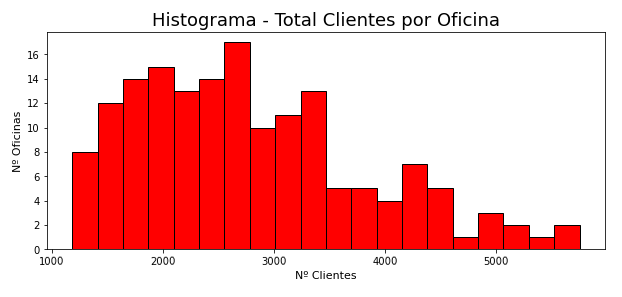
DZ nº8 11 oficinas. Corresponde a una DZ con pueblos.

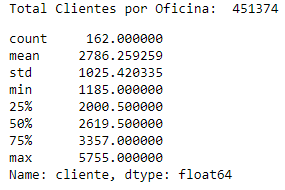


* Colectivo 451.000 Clientes.

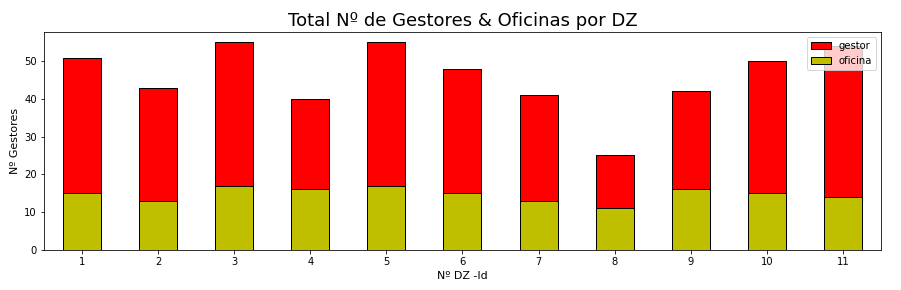
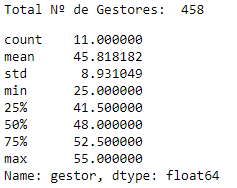
Grueso de oficinas entre 1500 y 3500 clientes

Oficinas con menos de 1500 clientes y mas de 5000 son escasas.





* Servicio Asesoramiento Especializado
* Colectivo de 458 Empleados
* No hay mucha descompensación entre DZs.
* Media de 48 Asesores – Gestores \*DZ8

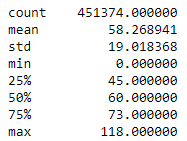


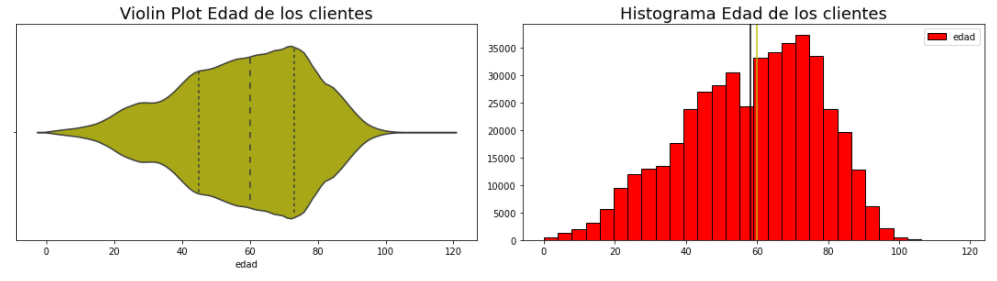
# ***Bloque 2 Edad de los clientes.***

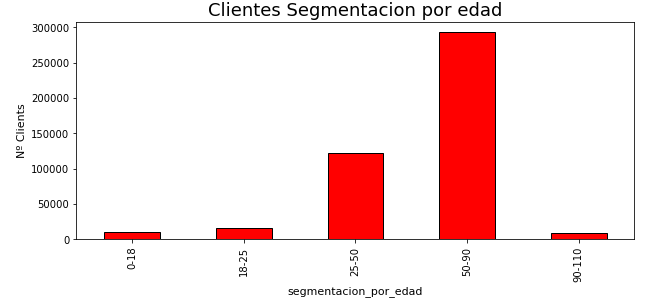
Visualizamos mediante un Box Plot la variable edad de los clientes. Esta visualización nos da la oportunidad de identificar la edad minina de los clientes, edad máxima, media, así como los cuartiles 25% y 75%.

Adicionalmente realizamos una segmentación de clientes por edad para ver en que rango de edad hay más clientes.

Las gráficas nos dicen que la mayoría de clientes se encuentra en un rango de edad de 50 a 90 años. Se trata de una base de datos con clientes de mediana y avanzada edad.

* La edad mínima que reflejan los datos son 0 años. (Valor incluido en estudio Nans)
* La edad máxima son 118 años.
* La media de edad está en 58.2 años.



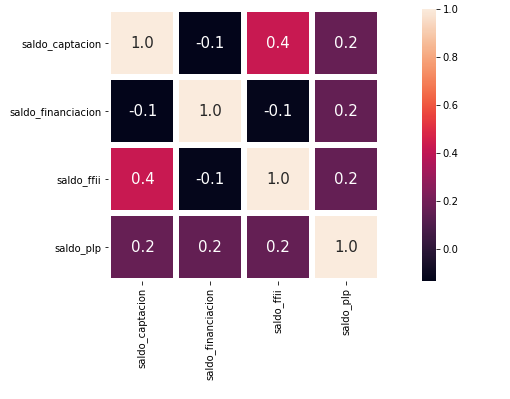


De cara a nuestro modelo, además de no resultar significativos por número total eliminaremos a los clientes entre los rangos de edad de 0 a 25 años y de 90 en adelante. La realidad a la hora de contratar un seguro de hogar es que los menores y jóvenes hasta los 25 años no contratan un seguro porque sencillamente no tienen vivienda propia. En cuanto a los clientes mayores de 90 años, dada su avanzada edad, normalmente son sus hijos los responsables de los seguros de hogar, por lo que vamos a eliminarlos del modelo.

# ***Bloque 3 Variables de Ahorro y Financiación*.**

Dentro de este bloque estudiamos las variables saldo captación, saldo financiación , saldos en fondos de inversión y saldos en plantes de pensión .

Iniciamos el estudio de este bloque mediante la visualización de posibles correlaciones entre las propias variables por si pudiésemos eliminar alguna de ellas por alta correlación.

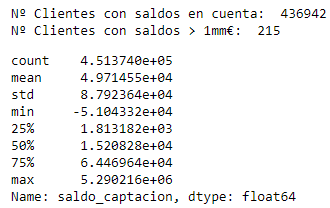
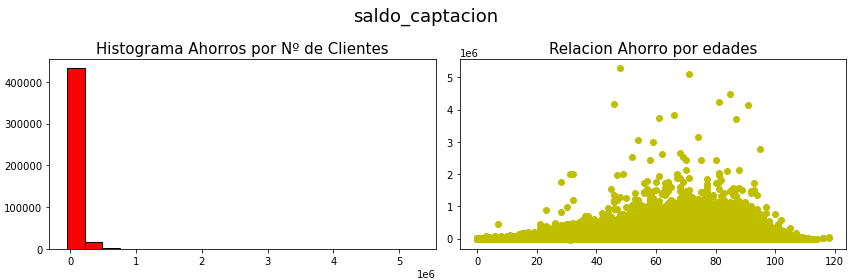


La grafica de correlaciones muestra que no existe grandes relaciones por lo que, a falta del estudio de cada una de las variables de forma independiente, en principio se serán mantenidas todas ellas de cara al modelo de clasificación.

Vemos una a una cada variable.

* Saldos de ahorro.

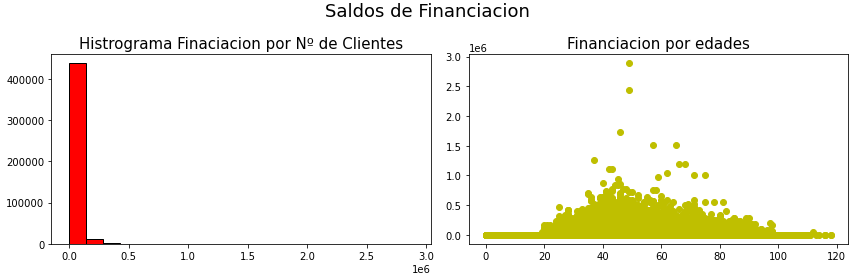
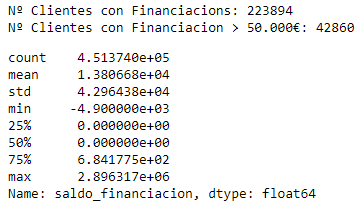
Conseguimos entender cuál es la realidad de los ahorros de los clientes mediante la visualización por histogramas y Scatter Plot, y mediante descripción estadística de la variable. Según arroja la función describe() la media de saldo en cuenta de los clientes es de menos de 5.000€. Si bien existen clientes con saldos importantes situándose el máximo en 5.2mm€ el 75% de los clientes no llega a los 6.500€ y los ahorros se distribuyen claramente entre los clientes a partir de 60 años hasta los 90. Solo 215 clientes disponen de mas de 1mm€ en cuenta.



* Saldos Financiación.

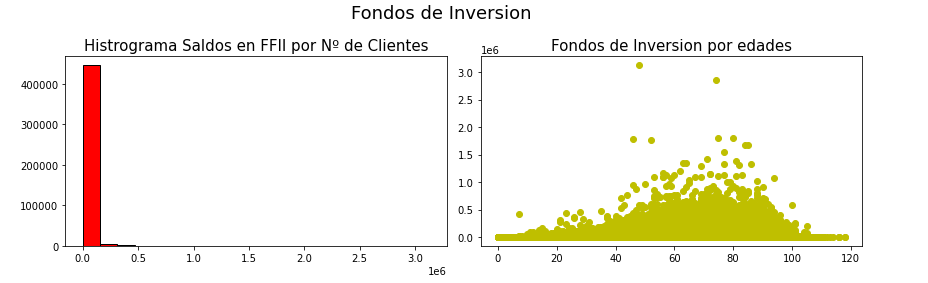
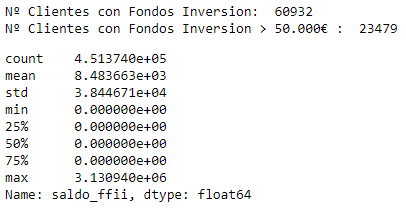
La mitad de los clientes de nuestra base de datos tienen prestamos, pero estos son de pequeño importe. Solo el 10% de los clientes tienen prestamos superiores a 50.000€, cantidades que pueden dar a pensar que la financiación corresponde a préstamos hipotecarios de adquisición de vivienda. El préstamo medio es de 13.000€ seguramente prestamos destinados a consumos familiares, vehículos, reformas, etc. El rango de edad donde mas prestamos tienen son de 30 a 75 años, edades que casan con la realidad del rango de edad en los que se solicitan préstamos al consumo. Por último, vemos que hay clientes con deudas.

Se trata de una variable importante que mantendremos en el modelo. Si bien daremos valor 0 a los clientes que tienen saldo negativo / deudas para no distorsionar.



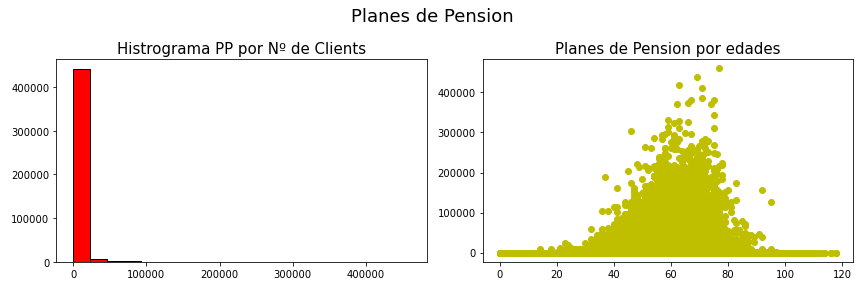
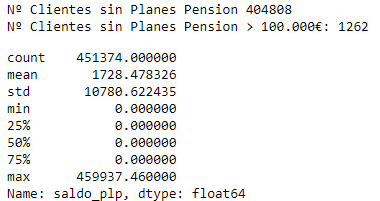
* Fondos de Inversión.

Se trata de una variable que mide el nivel de ahorro de los clientes y también de cultura financiera. No todos los clientes son susceptibles de tener este tipo de ahorro. Bien por desconfianza o por desconocimiento de su funcionamiento. Aquellos clientes que si disponen de Fondos de Inversión están dentro de un rango de edad de 50 a 90 años. Con la descripción estadística vemos claramente que la mayoría del cliente no tienen fondos. La media se sitúa en 8.000€.



* Planes de Pensión.

Se trata de un producto dirigido a la jubilación. Se trata de ahorrar durante la vida activa para cuando se llegue a la jubilación se pueda complementar la pensión que se recibirá de la seguridad social. En la misma línea de los fondos de inversión, se trata de un producto que no todo el mundo tiene a pesar de ser muy buena opción de ahorro aunque no la única. Los clientes que tienen planes de pensión se sitúan en rango de edad de 40 años hasta llegar al máximo con 65 años donde empieza a descender motivado por los rescates de los planes una vez jubilados los clientes.



La realidad comercial es que el hecho de tener saldos en Fondos de Inversión o Planes de Pensión no es una de las condiciones que llevan a los clientes a contratar o no un Seguro de Hogar. Estas dos variables se convertirán en variables de categoría booleana. Los valores asignados serán 0 si no se tiene y 1 si se tiene. De esta forma simplificaremos el estudio.

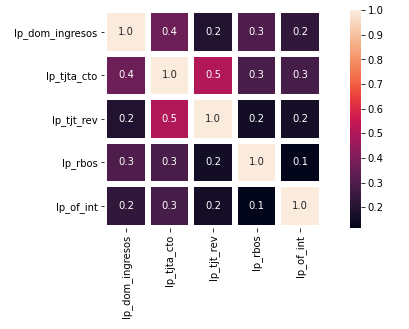
En cuanto a la variable de saldo en cuenta. Son valores reales y si pueden tener efecto en la contratación del seguro. Sólo modificaremos los negativos y les daremos el valor de la media.

Por último, vemos que el 50% de los clientes sí tienen financiación. Mantenemos la variable que puede ser útil para el modelo. A los clientes con saldo deudor les damos un valor de 0€.

*#* ***Bloque 4 Servicios***

Dentro de este bloque estudiamos los distintos servicios que disponen los clientes. Nóminas, domiciliación de recibos, tarjetas de crédito y App internet.

Visualizamos en primer lugar las correlaciones que puedan tener entre sí.



La mayor correlación que existe 0.5 es de tarjeta de crédito con tarjeta revolvíng. Se trata de dos tarjetas de crédito, pero de distinto tipo de uso. La primera el tipo de pago es 100% a fin de mes, mientras que la tarjeta revolving es de pago fraccionado mediante cuota definida fija.

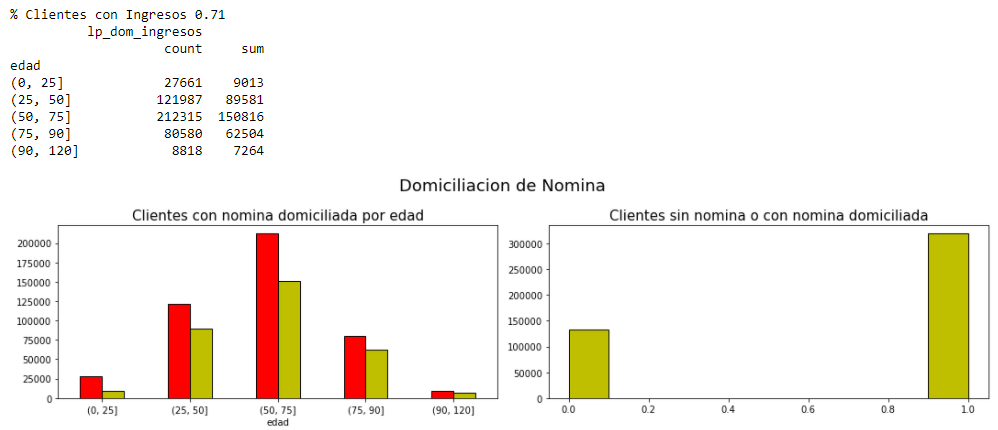
A pesar de esta correlación del 0.5 ambas tarjetas serán mantenidas en nuestro modelo ya que se trata claramente de dos realidades distintas.

Estudiamos una a una nuestras variables de servicio.

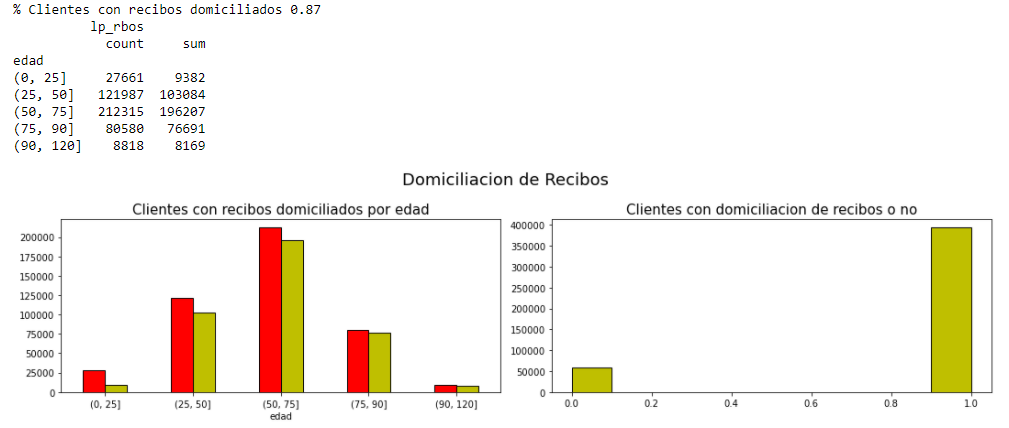
* Domiciliación de Nomina.

Variable muy importante. Quizás el servicio más demandado por una entidad financiera. La domiciliación de nómina hace referencia a la nómina del trabajo y a la nómina por pensión. El hecho de tener la nomina domiciliada en un banco supone una serie de beneficios mutuos para consumidor y entidad. El consumidor puede obtener mejores precios de préstamos, menos comisiones y mejores primas en seguros. Por otro lado, la entidad financiera a pesar de cobrar menos comisión a este tipo de clientes se beneficia en la propia comisión generada por la comercialización de los seguros y de servicios asociados que pueden conllevar..

Tenemos a un amplísimo porcentaje de clientes que tienen la nomina domiciliada y entre todo tipo de segmento de edad. Si bien se agudiza en edades de mas de 25 años manteniéndose hasta la propia jubilación con la domiciliación de la pensión.

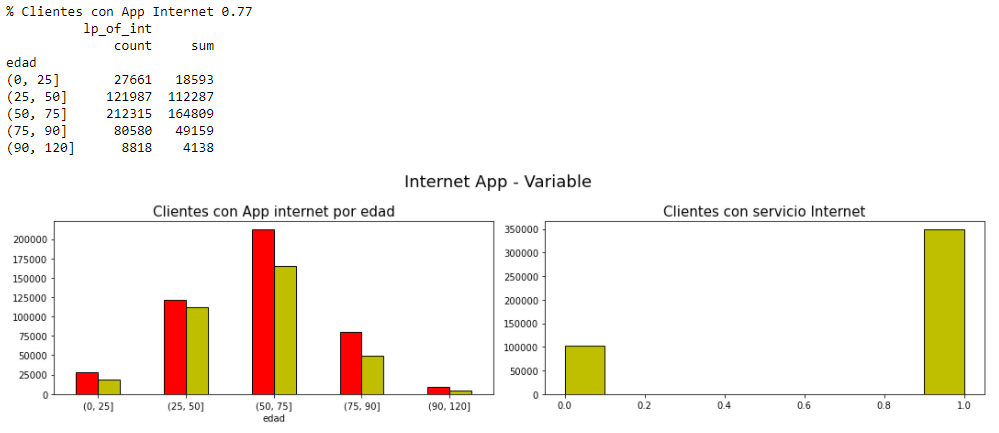


* Domiciliación de Recibos.

Este tipo de servicio suele ir de la mano a la domiciliación de la nomina . Esto se ve en las siguientes graficas que son muy parecidas a las del servicio nómina. El 87% de los clientes tienen el servicio. Por supuesto este tipo de servicio también deja importantes beneficios a un banco ya que para tener domiciliado un recibo es necesario disponer de saldo en cuenta y el saldo en cuenta es sencillamente margen de beneficio para la entidad.

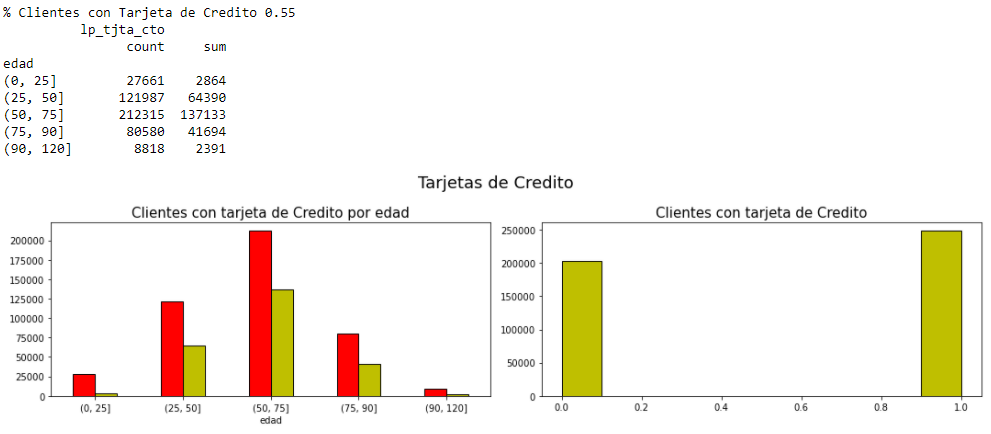
* App Internet.

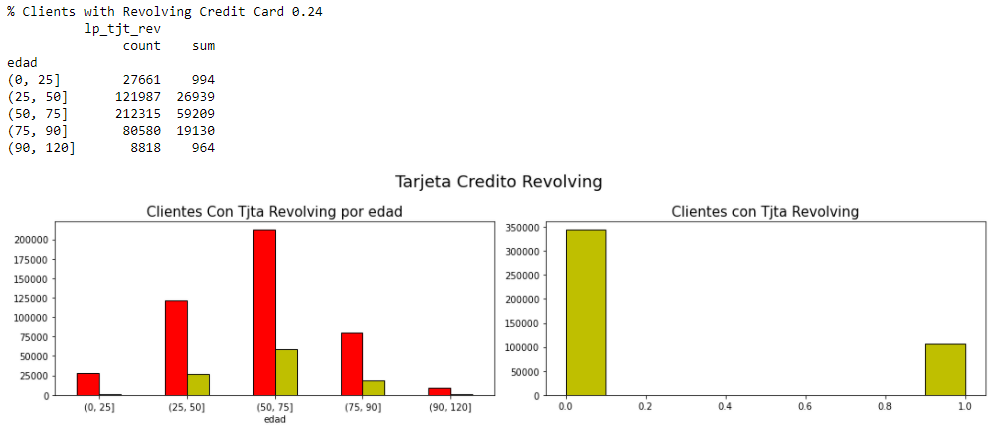
Servicio destinado a la consulta y operativa 24 horas de las posiciones de un cliente. A través de este servicio es posible realizar compras, transacciones y consultas sin que la oficina del banco este abierta. Es un servicio que tienen la mayoría de los clientes, el 77% en este caso. Los beneficios de este servicio son para ambos lados. Cliente puede realizar en cualquier momento una operación y la entidad financiera ahorra costes de estructura y esfuerzos logísticos.



* Tarjetas de crédito y Revolving.

Servicio que mantienen el 55% de los clientes en el caso de Tjta e Crédito pago 100% y el 24 % de los clientes en el caso de Revolving.



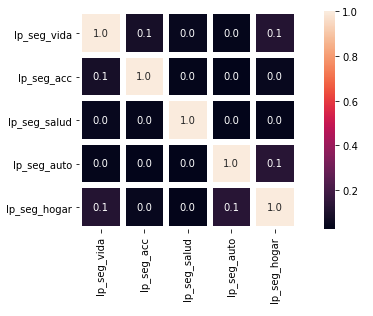


Como conclusión de este bloque n 4 # Servicios, podemos asegurar que el data set esta formado por clientes con media - alta vinculación con la entidad financiera. Aun asi, se evidencia un amplio potencial de desarrollo comercial. Las cinco variables que componen este bloque van casi de la mano tanto en tenencia como en el segmento de edad por utilización. Son clientes entre 25 y 75 años los que mas utilizan estos servicios.

Todas las variables serán mantenidas para el modelo ya que consideramos que son útiles para estimar si el cliente puede contratar el seguro de hogar.

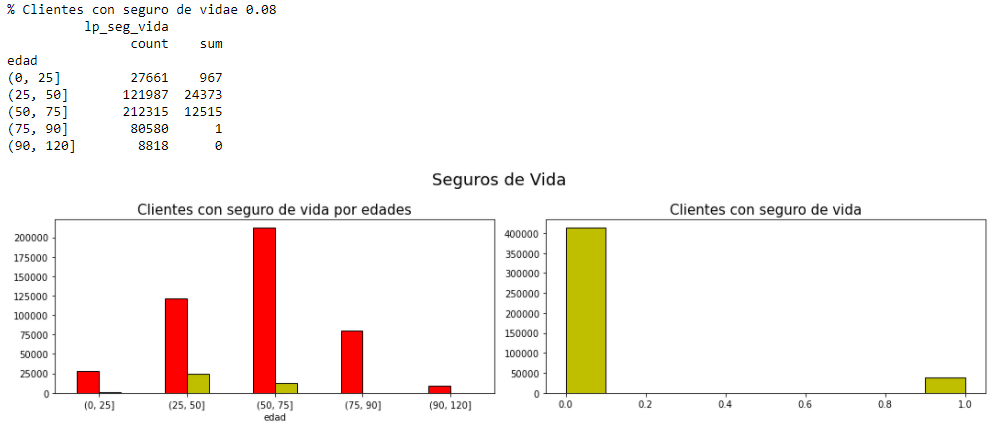
# ***Bloque 5 Seguros de Riesgo***

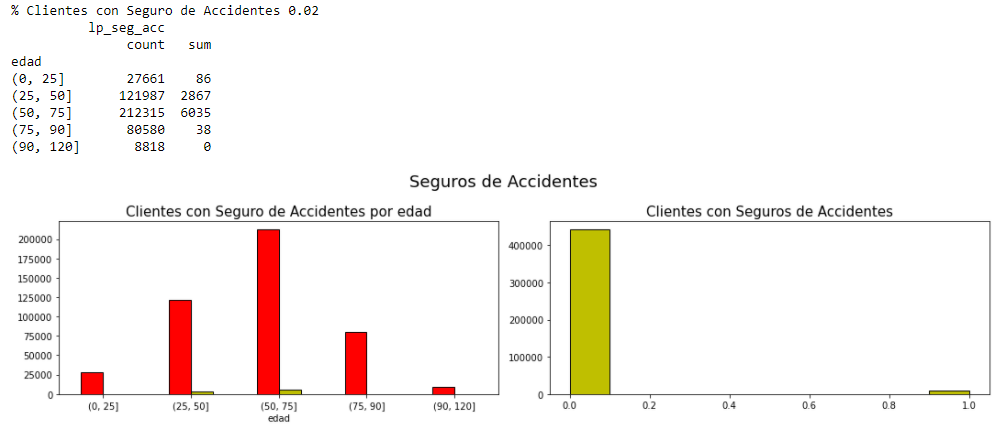
En primer lugar, vemos como las variables no tienen ninguna correlación entre sí.



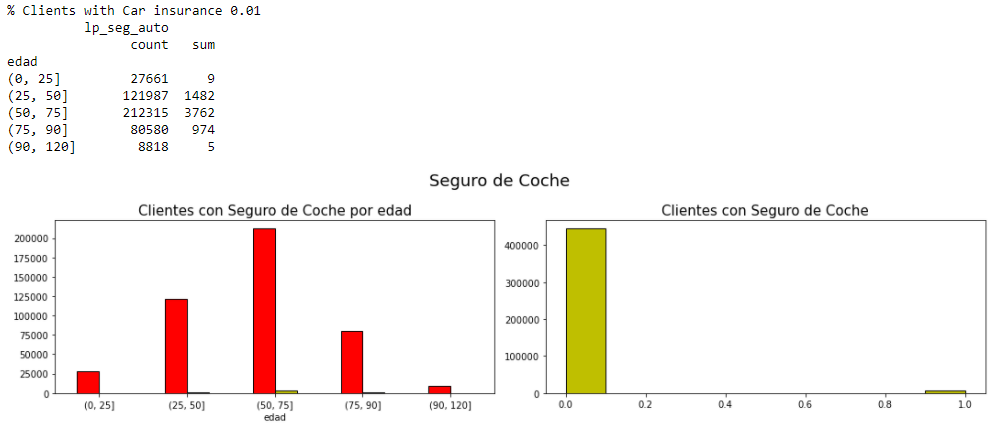
* Seguros de Vida, Accidentes, Salud y Auto.

Comprobaremos como estos seguros que atienden a contingencias de vida, accidentes y salud no tienen una buena penetración en los clientes. El seguro más contratado por los clientes es el seguro de vida y solo se trata de un 8%. Los porcentajes de tenencia de los otros seguros descienden hasta sencillamente el 0% en el caso de Salud. Anteriormente en el bloque de servicios hemos comentado como el hecho de tener nómina y recibos domiciliados puede ser una ventaja para los clientes por ahorrarse y beneficiarse de bajadas de prima en seguros. Estos datos reflejan el altísimo potencial comercial y de mejora en este segmento. Se trata de una oportunidad importante de generación de comisiones que no hay que dejar pasar. Estos datos cruzados con el servicio de recibos pueden ser importante ya que es posible detectar clientes con recibos de seguros en otras aseguradoras. Esto es una verdadera oportunidad comercial ya que la realidad es que en cualquier casa o familia lo normal es que haya uno o dos vehículos, el cabeza de familia tenga un seguro de vida o la familia tenga un seguro de salud. Este porcentaje tan bajo de posesión de seguros indica que probablemente los tengan en otra entidad financiera o compañía de seguros.







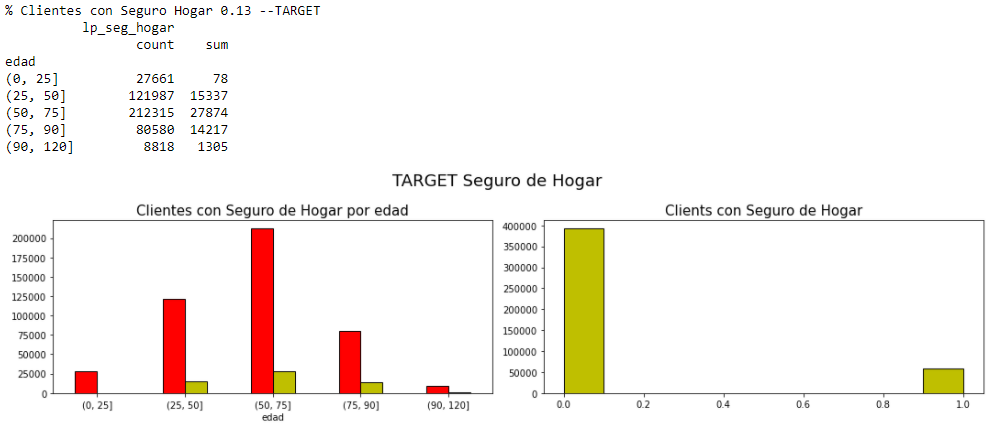


* Seguro de Hogar **TARGET.**

Estudiamos mas en profundidad nuestra Target. ¿Cuál es su estructura con respecto a nuestro marco general, que variables son importantes , podemos descartar alguna de ellas?

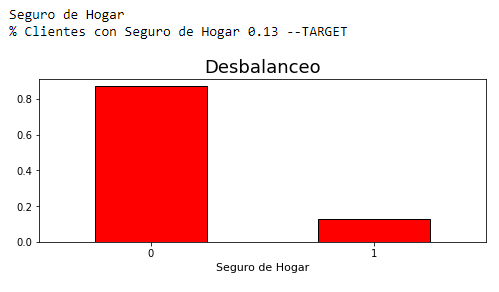
Visualizamos cuantos clientes tienen el seguro de hogar y como están distribuidos en nuestras direcciones de zona.

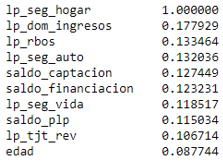
La DZ que más clientes tiene con seguro de hogar contratado es la DZ 3 con 7233 clientes mientras que la que menos tiene es la DZ 8 con 2.274 clientes. En todo caso, las cifras son bajas y la media se posiciona en 5.346 clientes. Estos datos vs un data set de 450.000 clientes reflejan un gran potencial de comercialización.

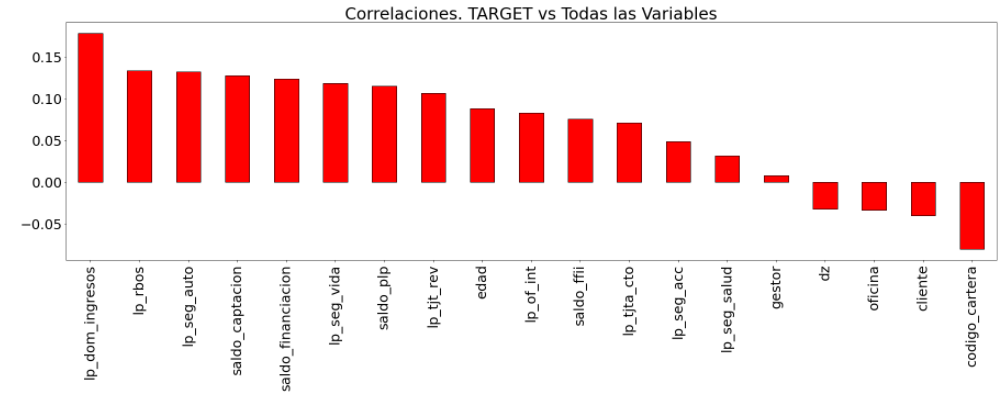


Vemos como es un seguro que tampoco tiene mucha penetración en los clientes. Solo el 13 % de los clientes de la base de datos tiene este seguro. El segmento de clientes con más contrataciones de seguros de hogar son entre 25 y 90. Ninguno de los extremos tienen contratacion de seguros relevantes. En este caso, atendiendo a estas magnitudes y confirmando lo estudiado en la variable edad, eliminaremos de nuestro modelo a clientes que de 0 – 25 años y de 90 – 120 años.

Esta cifra refleja otra circunstancia importante y es el gran desbalanceo que tiene. Esto tendremos que tenerlo en cuenta en la construcción del modelo.



¿Cuál es la correlación que tiene el seguro de hogar con todas las otras variables de la data set.? Vemos como aquellos que tienen nomina domiciliada y recibos domiciliados son los clientes que más relación tienen con el seguro de hogar. Aun así, la mayor correlación es de 0.17.



# ***Bloque 6 Variables de Segmentación***

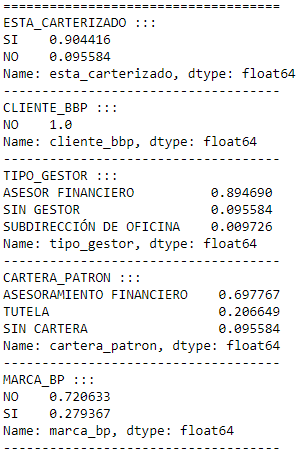
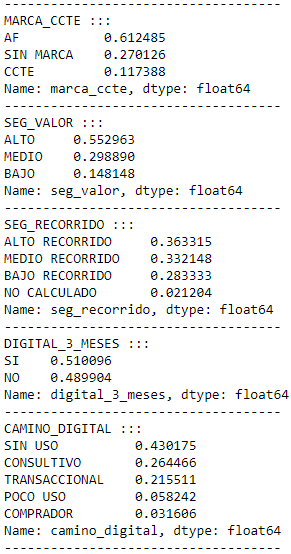
# ***Bloque 6 Segmentación***

Identificamos distribución de los clientes por segmentación

Estas variables hacen referencia a distintos ámbitos en los que un cliente puede estar segmentado. El 89% de los clientes están carterizados y asignado a un asesor financiero, aunque también hay clientes asignados a otras figuras como el subdirector.

Una variable importante es el segmento de recorrido. Nos indica que el 70% de los clientes tiene recorrido comercial. (No hace más que confirmar lo visto hasta ahora.)

Camino digital y Digital 3 Meses. Otras variables importantes que reflejan la poca utilización del servicio internet . El hecho de no utilizar este servicio esta asociado a la edad de los clientes. Además de ser un servicio que no tienen muchos clientes el segmento de edad donde mas se utiliza es de 25 a 50. Clara evidencia de la realidad tecnológica y de la falta de transformación y adaptación digital de los clientes de mediana y avanzada edad.



Variables categóricas “Esta Carterizado” y “BBP” serán eliminadas del modelo ya que el data set esta justamente construido sobre la base de un conjunto de clientes que tienen una cartera definida. Por otro lado, ningún cliente es BBP.

El resto de variables serán mantenidas a falta de un estudio mayor de importancia de las variables.

* 1. **Importancia de las Variables.**

Estudiamos la importancia de las variables utilizando un modelo de Random Forest y una matriz de correlaciones.

Con este estudio pretendemos revelar que importancia tiene cada una de las variables dentro de nuestro data set. Los datos revelados son los siguientes:

En cuanto a la matriz de correlaciones hemos utilizado la tabla de correlación de Spearman, cuyo resultado arroja correlaciones más altas que los demás métodos.

La correlación entre las variables es generalmente baja. Incluso hay correlaciones negativas, pero también bajas.

Basándonos en el resultado obtenido a través del modelo de Random Forest utilizado para revelar la importancia de cada una de las variables, vamos a eliminar del modelo por su escasa importancia las siguientes:

18) digital\_3\_meses 0,007651

19) lp\_rbos 0,007175

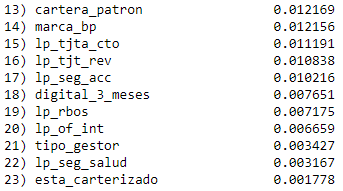
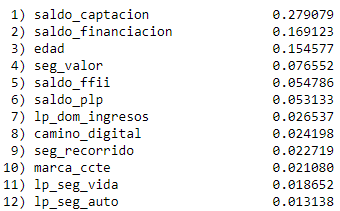
20) lp\_de\_int 0,006659

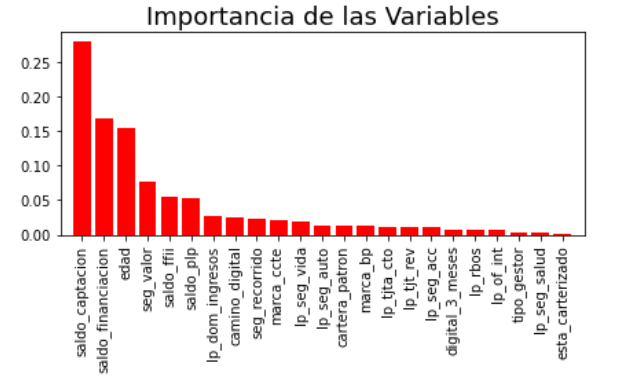
21) Tipo de gestor 0.003427

22) lp\_seg salud 0,003167

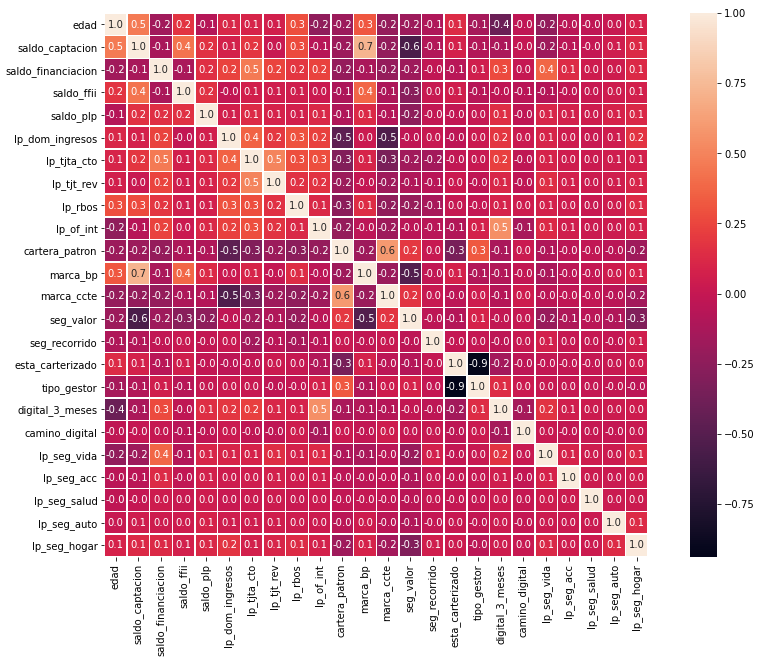
23) Esta\_Carterizado 0,001778. En este caso se confirma nuestra decisión anterior.

13) cartera\_patron 0,012169 Correlación con variable 10) marca\_ccte 0,021080





**Matriz de Correlaciones**



* 1. **Preprocesado.**

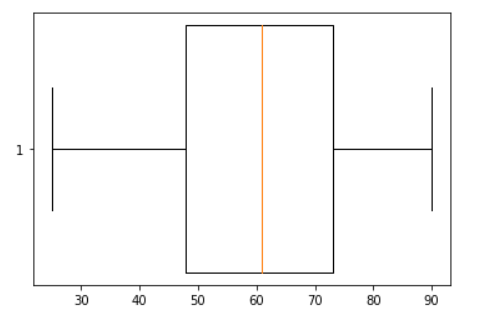
Recordamos que la preparación de los datos y el preprocesado de la base de datos comenzó inicialmente en el punto 4.6 Unión de archivos y Preparación de los Datos, justo después de haber construido y unido todos los datos en un solo archivo. En esa ocasión vimos y analizamos los Nans y les dimos valores acordes a las categorías de cada variable.

A lo largo del estudio EDA realizado, se han llegado a una serie de conclusiones las cuales vamos a poner en práctica en este capítulo. Realizaremos por lo tantos aquellas modificaciones necesarias para dejar limpio y preparado nuestra base de datos para la construcción del modelo.

**1º** Eliminamos 2 variables categóricas de segmentación. Está carterizado y cliente BBP.

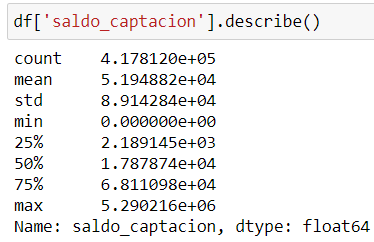
**2º** Estudio de Outliers . La única variable que puede tener Ouliers es la variable Edad. Al eliminar a los clientes entre los rangos de edad 0-25 y 90 según hemos concluido en el Eda confirmamos mediante grafica boxplot que finalmente la variable edad no contiene outliers.

Box Olot Variable Edad



**3º** Fondos de Inversión y Planes de Pensión son convertidas en variables booleanas. (0-1) Tiene o no tiene.

**4º** Variable Captación: Damos valor cero a todos aquellos valores negativos quedando conformado de la siguiente forma:

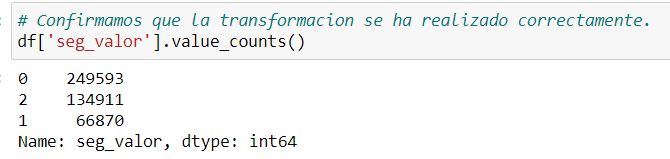


**5º** Label Encoder - Dentro de nuestras variables de segmentación que definen una condición especifica de cada uno de los clientes, encontramos finalmente cinco variables categóricas que queremos utilizar y que necesitamos convertir en número.

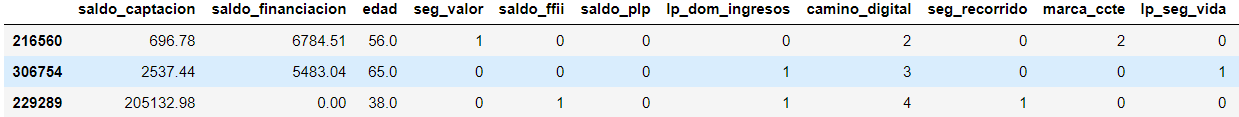


Sustituimos así cada uno de los posibles valores de definición por un número . Conseguimos mediante este método poder utilizar en nuestro modelo y en un mismo espacio todas las variables numéricas y categóricas.

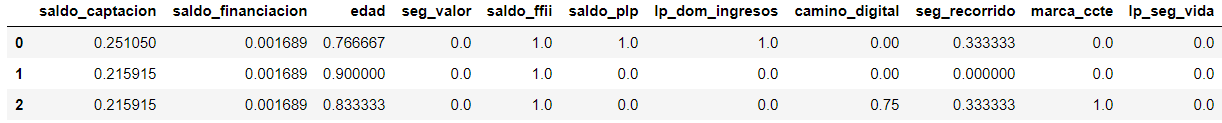
Antes de Label Encoder Después de Rabel Encoder



**6º** Normalización - Poco a poco vamos evolucionando nuestra base de datos. De las 17 variables definitivas que hemos dejado para el estudio del modelo, encontramos dos de ellas “Saldo Captación” y “Saldo financiación” con valores reales establecidos en Euros. Valores especificados en miles, cientos e incluso millones de euros. También tenemos la variable edad que debemos trabajar. Necesitamos convertir todas las variables a una misma escala común que no haga distorsionar o dar mayor importancia a unas variables numéricas de otras. Utilizamos el método de Normalización MinMaxEscaler y lo que conseguimos es traducir todas las variables a un rango entre 0 y 1

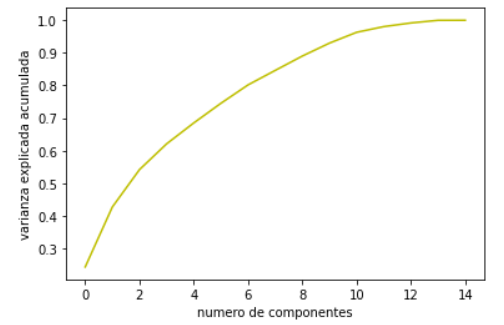
Antes de Normalizar

Después de Normalizar



**7º** PCA Posible simplificación de variables reduciendo dimensionalidad a la base de datos.

Después de haber reducido el Data Frame eliminando unidades de negocio y otras variables según estudio EDA, nos hemos quedado únicamente 17 variables. No consideramos necesario utilizar reducción de dimensionalidad PCA y lo confirmamos con la siguiente grafica en la que vemos como después del estudio de PCA a partir de 13 – 14 variables explicarían el 99 % de los 17 originales. Por lo tanto, utilizar PCA no implicaría una mejora sustancial del resultado.



**Conclusiones Previas a la Construcción del Modelo.**

Finalizamos con este capítulo la preparación, preprocesado de los datos para iniciar la construcción de nuestro modelo de clasificación. A lo largo de este proceso hemos generado hasta 3 base de datos distintas.

* Base de datos eliminando variables categóricas, cambiando a booleano variables de FFII y PP, eliminando clientes de ciertos rangos de edad y adaptando las variables categóricas con Label Encoder.
* Base de datos anterior a la que se le une la normalización a través de minmaxscaler para que todas las variables estuviesen en una misma escala.
* Base de datos anterior implementando PCA reducción de dimensionalidad.

Hemos modificado valores Nans, hemos eliminado outliers en aquella variable susceptible de tenerla, se han adaptado variables de captación y financiación eliminando valores negativos, se han adaptado variables de fondos de inversión y planes de pensión dándoles valores booleanos 0 – 1 y se han aplicado técnicas de reducción de dimensionalidad y de transformación de variables categóricas a numeral.

Con toda esta cirugía aplicada a la base de datos construida inicialmente, desarrollaremos el modelo de clasificación cuya finalidad recordamos es “Generar un modelo predictivo de clasificación que ayude a toda la fuerza comercial de las sucursales a orientar la comercialización, a optimizar los tiempos, metodologías y sistemas utilizados. Todo ello en busca de un mayor éxito de ventas , margen para la entidad financiera y satisfacción de los clientes”

Todo el proceso de unión está ubicado en el notebook “preprocessing” al cual se puede acceder dentro de la carpeta Preprocessing del repositorio.

**Construcción del Modelo.**

Recordamos cual era nuestro Objetivo y Finalidad:

**Objetivo:** Desarrollar un modelo predictivo de compra de productos financieros concretando en los Seguros del Hogar.

**Finalidad:** Generar un modelo predictivo que ayude a toda la fuerza comercial de las sucursales a orientar la comercialización, a optimizar los tiempos, metodologías y sistemas utilizados. Todo ello en busca de un mayor éxito de ventas y satisfacción de los clientes.

**Elección del Modelo:**

El modelo final de clasificación elegido ha sido RandomForest

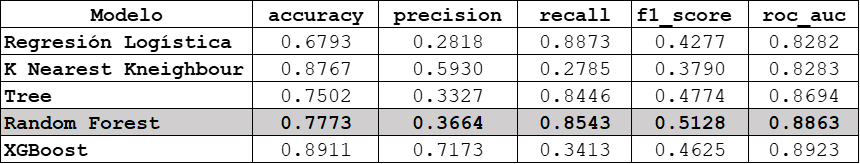
Buscamos una sistemática comercial que simplifique el estudio de los clientes antes de llamarles. Buscamos una optimización del tiempo para llegar a cuantos mas clientes mejor. Nos interesa finalmente filtrar aquellos clientes a los que tenemos que llamar. Este filtro se puede obtener desde dos perspectivas. Y lo buscamos dentro de una matriz de confusión. Buscando verdaderos positivos para cargarlos en el objetivo de contactos diarios y que los comerciales puedan llamarles o identificando verdaderos negativos para desecharlos y finalmente coger los verdaderos positivos y llamarles. En ambos casos llegamos a la misma conclusión y obtenemos lo que realmente queremos. Llamar a los que nos interesan.

Se han desarrollado 5 modelos distintos de clasificación con el fin de buscar aquel que mejor métricas y mejores comportamientos tenga según nuestras necesidades concretas. Las principales métricas que utilizaremos como evaluadores de nuestros modelos son Recall y Roc Auc.

**Modelos Machine Learning Entrenados.**

* Modelo Clasificación Regresión Logística
* Modelo Clasificación K-Nearest Neighbor
* Modelo Clasificación TREE
* Modelo Clasificación Xgboost
* Modelo Clasificación Randon Forest - MODELO FINAL ELEGIDO

Mostramos de inicio y a efectos comparativos una Tabla de métricas obtenidas de cada uno de los modelos entrenados.



El desarrollo y código de todos estos modelos se encuentran dentro de la carpeta “ Models“ de este repositorio.

Como hemos comentado al inicio, en esta memoria vamos a centrarnos en el **Modelo Final Elegido RANDOMFOREST.** Mostraremos como iniciamos el modelo, nuestra primera aproximación, como se ha evolucionado y desarrollado, problemas encontrados y soluciones implementadas, evolución de métricas obtenidas y finalmente la conclusión y resultado final.

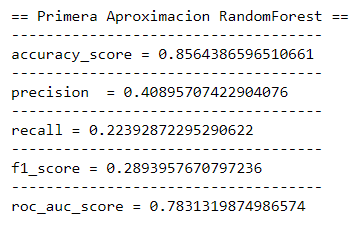
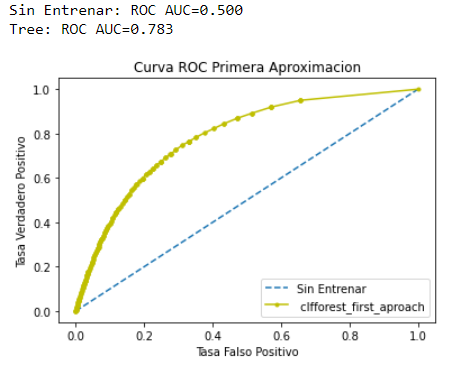
El motivo de la utilización de este modelo es principalmente por su sencillez de explicación su sencillez de funcionamiento y la comparativa positiva en métricas con otros modelos.

Se trata de un modelo fácil de interpretar muy útil para un modelo de clasificación, reduce de por si la dimensionalidad de las variables .

* **Primera Aproximación**

Nuestro camino comienza con la base de datos limpia con un preprocesado único de modificación de Nans. Se trata de la primera base de datos obtenida después de haberla unido en la primera fase de construcción. Queríamos saber con los datos en crudo que comportamiento tenia el modelo y que métricas obteníamos sin hacer nada intermedio.

Los Resultados fueron los siguientes. Claramente mejorables.



* **Evolución del modelo.**

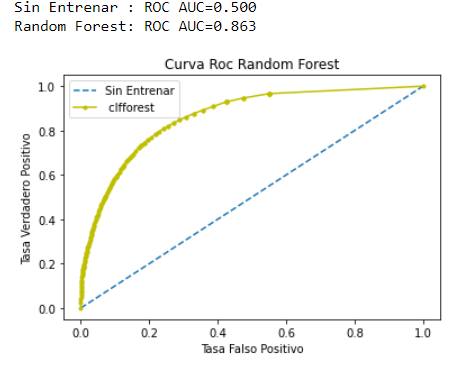
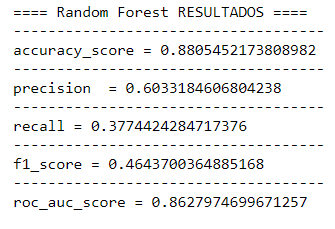
Comenzamos el estudio real del modelo. Recordamos que durante el preprocesado de la base de datos, se generaron hasta 3 bases de datos distintas.

1. Base de datos eliminando variables categóricas, cambiando a booleano variables de FFII y PP, eliminando clientes de ciertos rangos de edad y adaptando las variables categóricas con Label Encoder.
2. Base de datos anterior a la que se le une la normalización a través de minmaxscaler para que todas las variables estuviesen en una misma escala.
3. Base de datos anterior implementando PCA reducción de dimensionalidad.

El modelo ha sido entrenado con las dos primeras bases de datos. La tercera base con PCA implementada no la vimos necesaria dado que lo único que hacíamos era reducir dos variables.

El entrenamiento y resultado de las dos primeras bases de datos ha sido casi idéntico. Observamos que el hecho de haber normalizado los datos no ha sido un mejor condicionante para obtener mejores resultados. La explicación de esto es que en un modelo de RandomForest, la normalización no tiene por qué ser necesaria ya que no se comparan magnitudes. Se trata de dividir rangos y no de compararlos. Por otro lado, durante el preprocesado hemos visto como la importancia de las variables no difería mucho. Estos motivos han hecho que en el estudio final no se haya utilizado la base de datos normalizada.

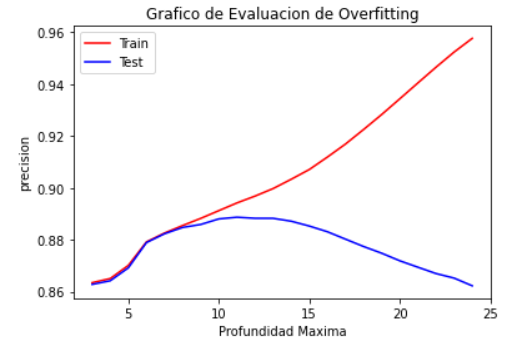
La primera decisión que tomamos en esta primera evolución del modelo es utilizar un estimador que intente poner remedio al gran desbalanceo de nuestro target. Utilizamos el estimador de “ Class Weigth = Balanced” . Las métricas obtenidas solo con este ajuste mejoran algo nuestra primera aproximación, pero en ningún caso lo suficiente.

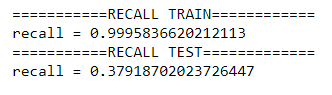


**Evaluación de Overfitting:** Lo que si evidenciamos en esta primera evolución del modelo es la existencia de Overfitting.

Los resultados obtenidos en train y test muestran de inicio que el modelo no está generalizando bien. Se obtienen métricas de 0. 99 en train y 0.37 en Test. Hay una grave situación de sobreajuste confirmada mediante Cross Validation, que tenemos que resolver.

A través de la siguiente grafica se muestra como el modelo de RandomForest, a partir de 8 profundidades empieza a no generealizar bien, evidenciándose el Overfitting



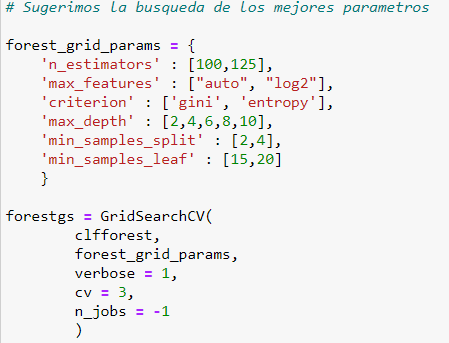


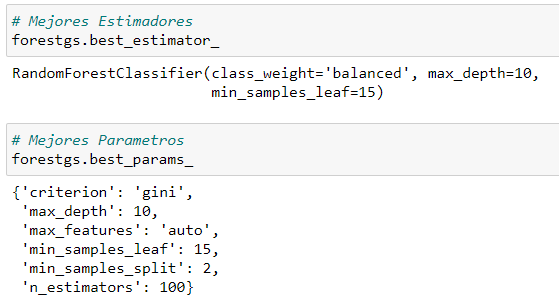
Nec

Podemos decir que este intento de mejora del modelo no es satisfactorio. Hemos confirmado que no se ha generalizado bien, que existe overfitting, que el desbalanceo del target no lo soluciona del todo y que las métricas requeridas no son acorde a nuestra necesidad. Conclusión. Hay que seguir evolucionado para mejorarlo.

Para ello **utilizaremos Gridsearch** que “permite evaluar y seleccionar de forma sistemática los parámetros de un modelo. Indicándole un modelo y los parámetros a probar, puede evaluar el rendimiento del primero en función de los segundos mediante validación cruzada”

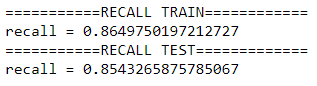
#Parámetros sugeridos:



#Parámetros recomendados mediante GridSearch:

* **Adaptación del Modelo con mejores parámetros recomendados:**

Implementamos los mejores parámetros recomendados.

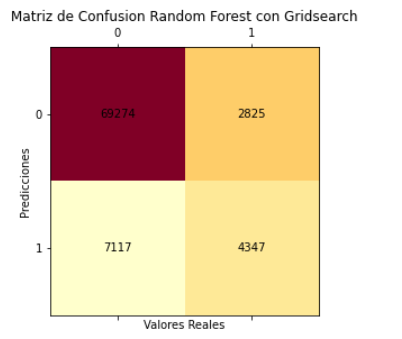
La primera mejora obtenida es que hemos conseguido solucionar el overfitting igualando las métricas de Train y Test a 0.85 Mediante Cross Validation se confirma. Primer problema solucionado.

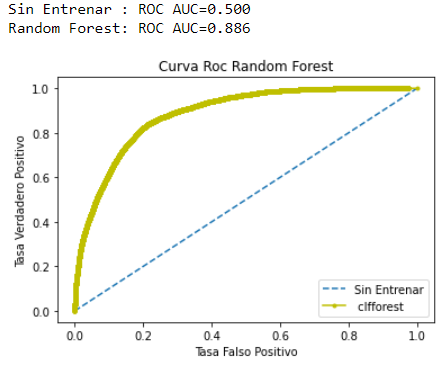


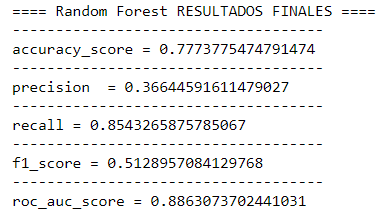


**Estudio de la Matriz de Confusión.**

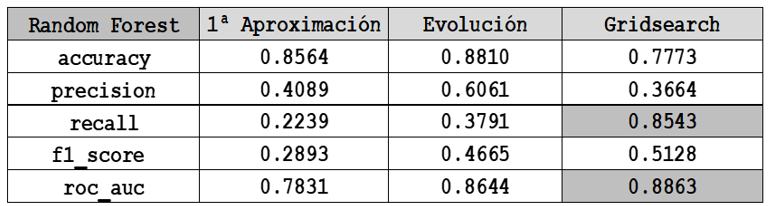
En la diagonal de aciertos obtenemos 69274 verdaderos positivos y 4347 verdaderos negativos. En nuestro modelo prima la detección de los clientes que sí quieren contratar el seguro. Necesitamos obtener una buena tasa de verdaderos positivos o verdaderos negativos. Si detectamos a los clientes que probablemente vayan a contratar o no un seguro de hogar, podremos trabajar de forma más directa. Todo ello con el objetivo de optimizar el tiempo y la eficiencia comercial. La realidad comercial es que da lo mismo trabajar con verdaderos positivos o negativos. Ambos sirven de filtro para poder llamar o no llamar. El resultado de Recall obtenido es óptimo, con un 85% y un 88% en AUC.



* **Resultados Finales y evolución de métricas**



* **Evolucion de Resultados:**

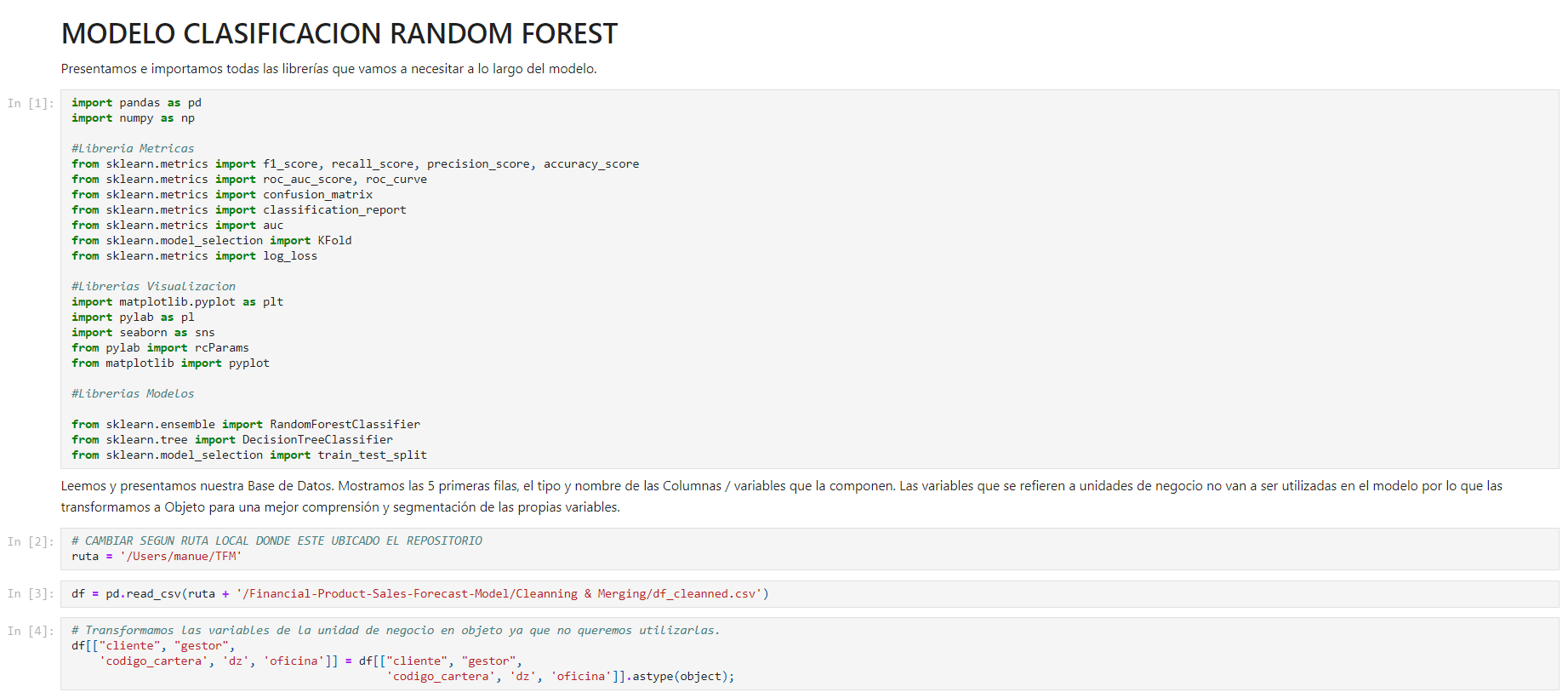


**5 Conclusiones del modelo.**

1. Las métricas elegidas\_\_ para evaluar nuestro modelo de clasificación Random Forest son Recall y AUC - Curva Roc.
2. Comenzamos nuestro modelo con una primera aproximación obteniendo resultados muy mejorables. Recall 0.22\_\_ La base de datos utilizada es la original sin ningún tipo de preprocesamiento.
3. Evolucionamos nuestro modelo. Utilizamos como base de datos df\_enocoded ya preprocesada en nuestro bloque de preprocesamiento. (Adaptado, aplicado Label Encoder y sin aplicar Standar Scaler ni PCA).
4. Primeros problemas. Desbalanceo y Evidencia de Overfitting.

Aplicamos estimador class weigth balanced para trabajar el desbalanceo de la base de datos. Por otro lado vemos cómo nuestro conjunto de entrenamiento no ha generalizado bien. Hay una gran diferencia entre los resultados obtenidos en Test y Train. 99 de entrenamiento - 37 de prueba. Podemos en este caso confirmar la existencia de overfitting. Todo esto lo podemos ver en una gráfica de ajuste del entrenamiento donde podemos ver que después de 8 nodos Train y Test empiezan a separarse - Probamos por Validación Cruzada obteniendo medias de diferentes métricas. Las métricas resultantes de esta primera evolución del modelo son 37% Recall y 86% Auc.

1. Dado que nuestro Target está muy desequilibrado con un 87% de No y un 13% de Sí, se ha utilizado de inicio el parámetro 'class\_weight = 'balanced' en la configuración del modelo, sin obtener mejoras evidentes.
2. Buscamos los mejores parámetros. Necesitamos mejorar el modelo y para ello utilizamos GridSearch para buscar soluciones . Una vez obtenido, volvemos a entrenar y obtenemos de nuevo las métricas Recall 92% Train - 91% Test para confirmar entre otras cosas que hemos conseguido solucionar el overfitting y que nuestro modelo ha conseguido generalizar bien. Confirmamos con la Validación Cruzada obteniendo las medias de las diferentes métricas con 5 validaciones.
3. Las métricas finalmente elegidas son Reccal 0,85 y AUC 0,88 que consideramos adecuadas para la explicación de nuestro modelo.
4. Matriz de Confusión. En la diagonal de aciertos tenemos 69264 y 4358. En nuestro modelo prima la detección de los clientes que sí quieren contratar el seguro. Necesitamos obtener una buena tasa de verdaderos positivos. Recordemos que nuestras métricas objetivo son Recall y AUC. Si detectamos a los clientes que probablemente vayan a contratar un seguro de hogar, podremos trabajar de forma más directa. Todo ello con el objetivo de optimizar el tiempo y la eficiencia comercial. El resultado de Recall obtenido es óptimo, con un 85% y un 88% en AUC.
5. Nuestra base de datos comenzó con 30 variables. Después de reprocesarlas y estudiarlas nuestro modelo se ha construido finalmente con 17 variables. Tenemos conocimiento que nos faltarían muchísimas variables adicionales que mejorarías seguramente el estudio. Entre ellas, variables y datos como los ingresos del cliente, los hijos, el tipo de trabajo, el salario real en euros obtenido, el lugar de residencia., compras realizadas, tipos de compra…etc..

## CODIGO MODELO ##

