StoryTelling

Participantes: Marcelo Baeza, Luis Salazar, Lucas Ferrero

# Introducción

Las tarjetas de crédito son un método común de control de riesgos en la industria financiera. Utiliza la información personal y los datos presentados por los solicitantes de tarjetas de crédito para predecir la probabilidad de futuros incumplimientos y préstamos de tarjetas de crédito. Así mismo, el banco puede decidir si emite una tarjeta de crédito al solicitante. En términos generales, los modelos se basan en el análisis de datos históricos y la formulación de modelos de predicción.

# Objetivo

Un banco desea identificar patrones en los clientes, lo que les ayudaría a tomar acciones sobre cuáles clientes serían los más apropiados para recibir una tarjeta de crédito.

Para esto, en base a datos proporcionados por el banco, se deberá generar un análisis para identificar los mejores patrones que respondan las preguntas sobre la necesidad de automatización de asignación de tarjetas de créditos.

# Descripción de la temática de los datos

En el presente análisis se procederá a trabajar con datos confidenciales asociados a clientes que poseen relación con el banco que solicitó el análisis

# Alcance

El alcance del proyecto está orientado a un banco que pretende evaluar y diferenciar a los clientes, ya sea por capacidad económica, capacidad de devolución, compromiso con las deudas, etc. Si bien este trabajo fue realizado con el dataset de un banco en especifico y no se puede difundir, otros bancos podrían usar el mismo modelos para realizar el análisis con sus propios clientes.

# Hipótesis

Al tener el historial de todos los clientes se puede analizar cuales son los clientes más confiables al momento de aceptar la solicitud de una tarjeta de crédito o al momento de darles un préstamo. Para esto el banco puede utilizar el historial de deudas de los clientes así como el tiempo que tardaron en pagarlas. Esto a la larga generará una mayor eficiencia y menor pérdida al banco.

# Herramientas tecnológicas implementadas

Para el presente trabajo se utilizaron los siguientes programas/herramientas:

1. Lenguaje de programación Python: permitió procesar la información inicial para poder generar las asociaciones entre toda esta.
   1. Librería que permiten operar sobre los datos
      1. Pandas
      2. Sckylearn
      3. Numpy
      4. Matplotlib
2. Google Colab, con el propósito de ejecutar las pruebas.

# Recorrido

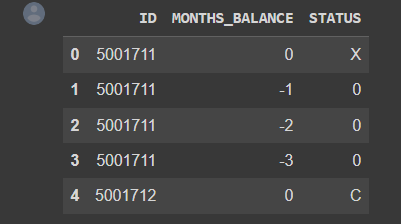
A continuación se expondrá el trabajo que se ha podido realizar para poder tener un análisis acorde a la necesidad del cliente

## Datos

Los datos con los que se trabajaron son los siguientes:

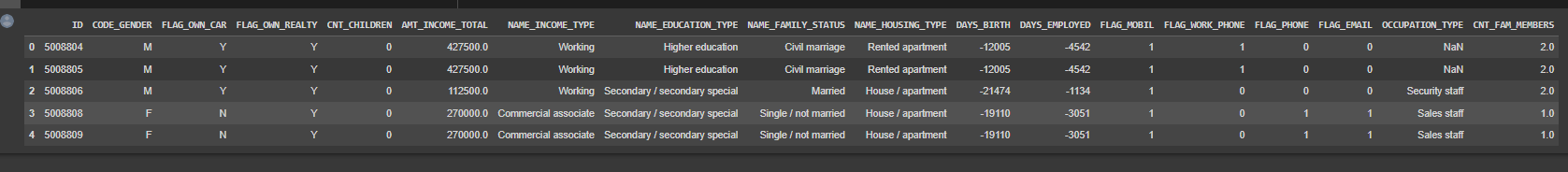
### Credit Record:

Información que indica si al cliente se le aprobó o no la tarjeta de crédito



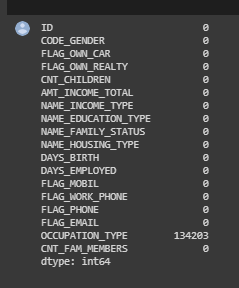
### Application Record:

Información que indica toda la información asociada al cliente



## Revisión de datos nulos

Se procedió a revisar qué datos son nulos. En ese caso la ocupación de la persona, interpretándose como que están desempleados o que no está registrado la información



## Criterios de asignación

Se definió el siguiente criterio para asignar una tarjeta de crédito.

*“un cliente será considerado como "malo" si ha tenido mora mayores o iguales 60 días en los últimos cinco años”*

## Análisis

### Conociendo los datos

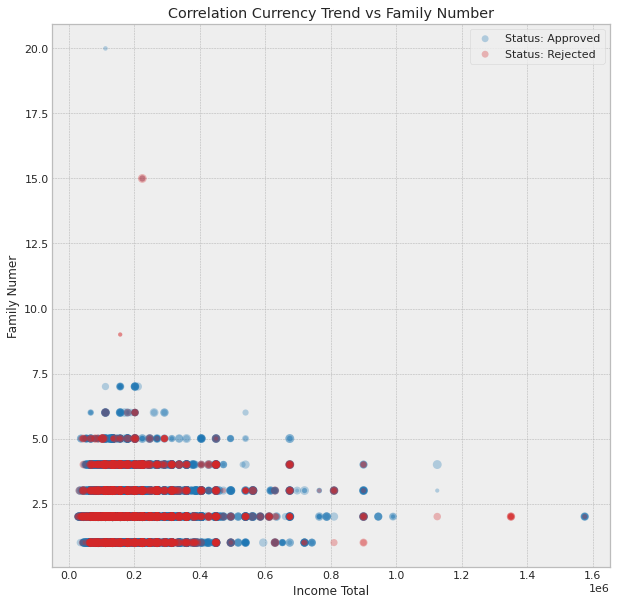
A continuación se exponen datos que permitieron familiarizarse con la información y generar estudios posteriores.

#### Relación de estado civil o aprobación de tarjeta de crédito

### 

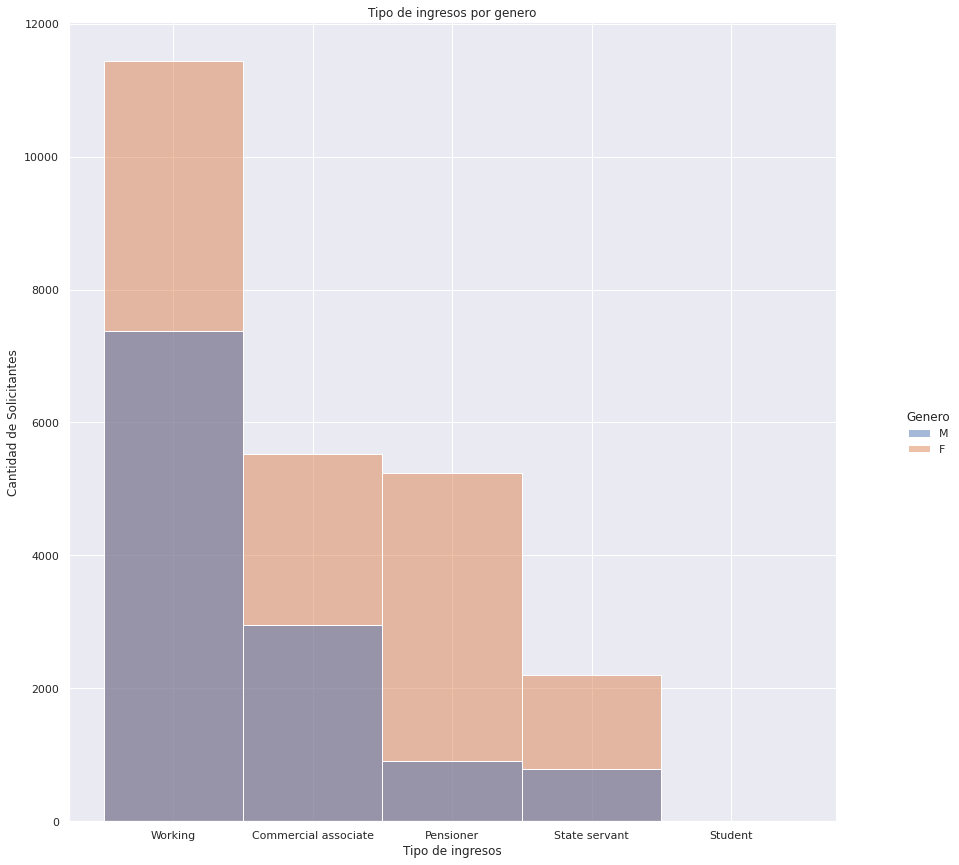
De este gráfico es posible interpretar que existe una gran afluencia de las personas que están casadas y que desean poseer una tarjeta de crédito; pero además no es posible concluir que existe alguna tendencia respecto al estado civil de las personas que indique que se le otorgará con mayor probabilidad su tarjeta de crédito

#### Relación de sueldo v/s el número de personas que viven en el hogar



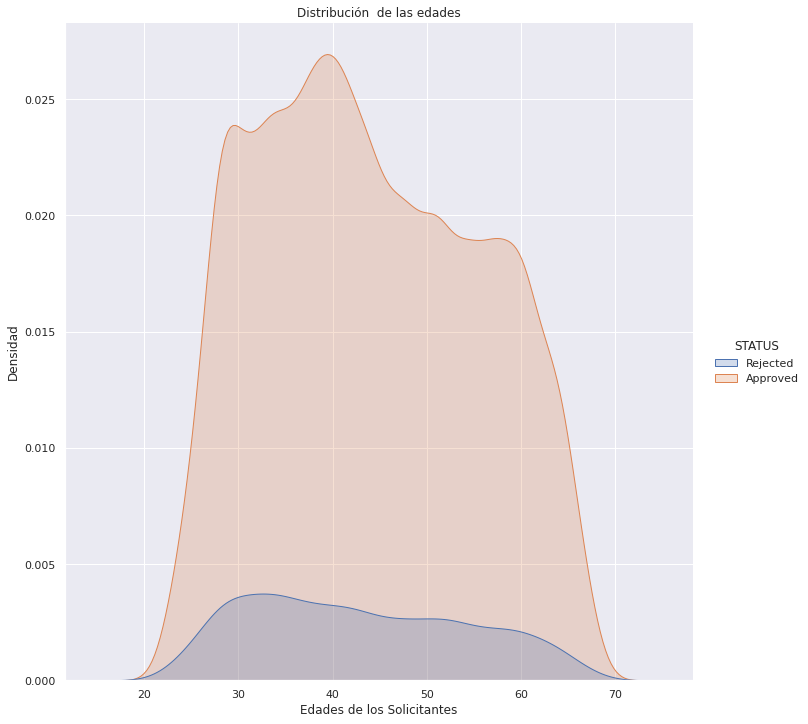
Respecto a la gráfica presentada, es posible concluir que la mayor cantidad de personas que solicitan una tarjeta de crédito tienen un salario anual máximo hasta los 500.000 USD. Además tampoco existe una relación directa entre los miembros familiares vs el salario. Solo se aprecia que dentro del rango especificado predomina el rechazo a la solicitud de tarjeta de crédito.

#### Tipo de ingreso por género



Con respecto al total de los solicitantes, se observa que la mayoría tiene sus ingresos en empleos privados en condición de dependencia, teniendo una mayor incidencia en este grupo los hombres. El segundo grupo corresponde a ingresos provenientes de negocios o inversiones, al igual que el caso anterior los hombres tienen mayor participación. Finalmente, están los pensionados y empleados públicos donde las mujeres tienen mayor incidencia.

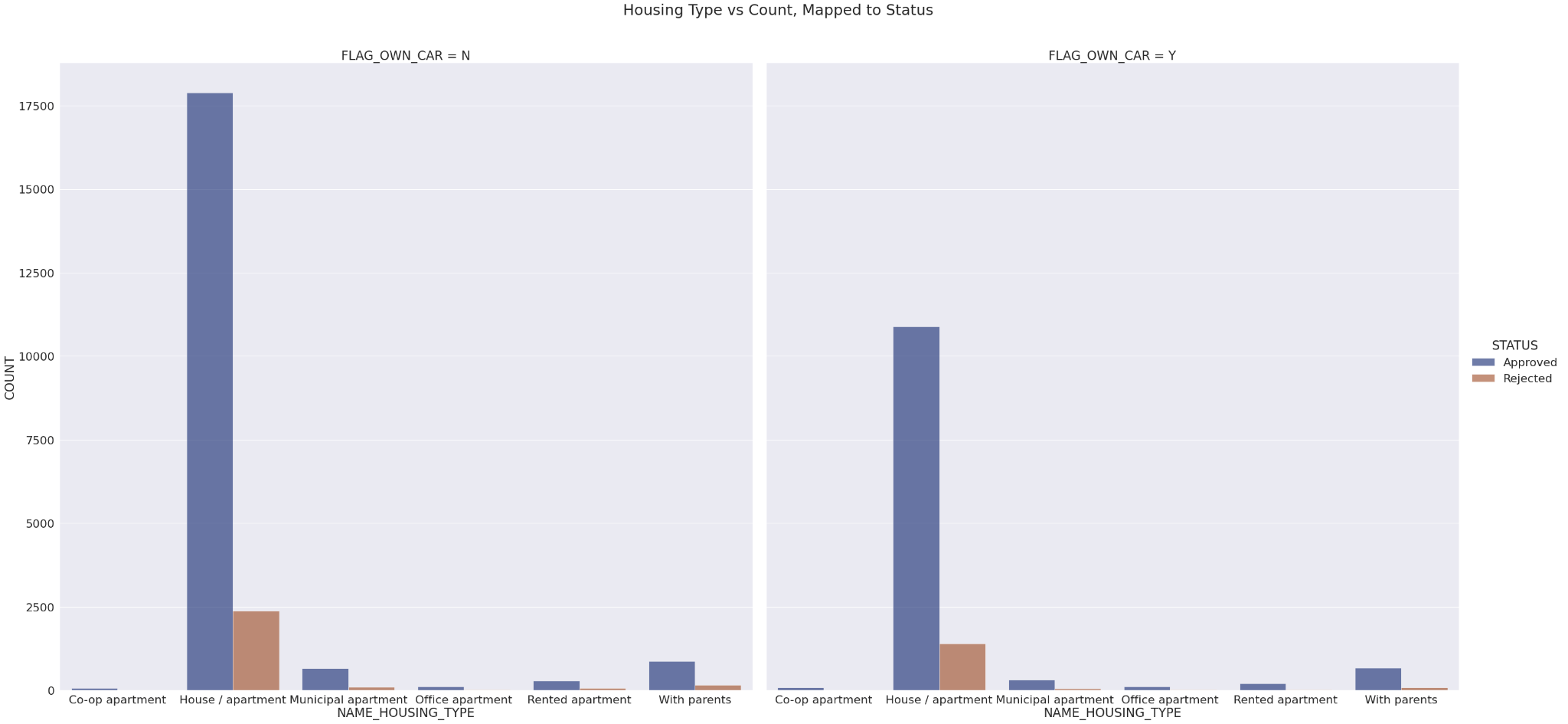
#### Distribución de Edades



Se observa que la mayoría de solicitantes y aprobados tiene edades entre 35 y 45 años. También es claro que ambas distribuciones tienen un comportamiento similar con lo cual se concluye que la edad no es factor que tuvo impacto en las aprobaciones. Finalmente, otro dato de relevancia es que la mayor parte de solicitantes y aprobados tiene edades por debajo de 50 años.

#### El tipo de hogar que viven los solicitantes

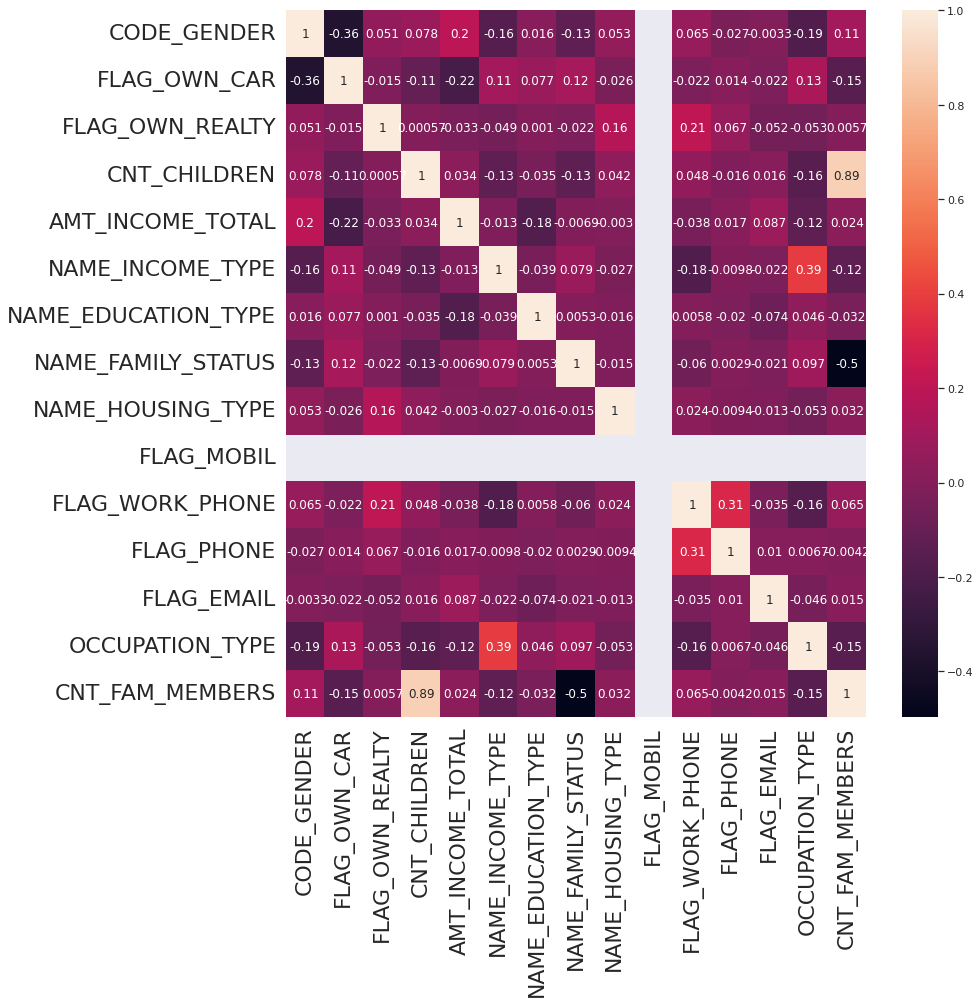
Esta información es para conocer el número de participantes que se les ha asignado una tarjeta de crédito, relacionada al tipo de “hogar” donde viven.



Para el análisis de la gráfica en donde se presentan dos escenarios: las personas que poseen o no un vehículo, y a la vez comparando su situación respecto al tipo de casa en que vive. La predominancia en ambas gráficas es para personas que poseen una casa o departamento. Además concluir que no existe una tendencia respecto a qué personas se le asigna una tarjeta o quienes no.

### Relación de las variables

Se expondrá las relaciones que existen entre las variables para conocer una tendencia entre los datos



Variables con mayor correlación:

* FLAG\_OWN\_CAR
* CNT\_CHILDREN
* AMT\_INCOME\_TOTAL
* NAME\_EDUCATION\_TYPE
* NAME\_FAMILY\_STATUS
* NAME\_HOUSING\_TYPE
* FLAG\_EMAIL

Respecto a la correlación trabajada, es posible concluir que solo existe una correlación positiva pero muy leve, ya que las variables que se indicaron son positivas, pero con un valor muy cercano a cero. Por lo tanto el grado de variación conjunta entre las variables es bajo.

##### Resumen de variables con aprobación a poseer una tarjeta de crédito

| FLAG\_OWN\_CAR informacion:  count 32166  unique 2  top N  freq 19892  Name: FLAG\_OWN\_CAR, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  NAME\_EDUCATION\_TYPE informacion:  count 32166  unique 5  top Secondary / secondary special  freq 21887  Name: NAME\_EDUCATION\_TYPE, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  NAME\_FAMILY\_STATUS informacion:  count 32166  unique 5  top Married  freq 22134  Name: NAME\_FAMILY\_STATUS, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  NAME\_HOUSING\_TYPE informacion:  count 32166  unique 6  top House / apartment  freq 28766  Name: NAME\_HOUSING\_TYPE, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  FLAG\_EMAIL informacion:  count 32166  unique 2  top Y  freq 29346  Name: FLAG\_EMAIL, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  CNT\_CHILDREN - Media: 0.4271902008331779  CNT\_CHILDREN - Desvación Standar: 0.7337435438029282  CNT\_CHILDREN - Moda: 0 0  dtype: int64  CNT\_CHILDREN - Mediana: 0.0  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  AMT\_INCOME\_TOTAL - Media: 185785.986149972  AMT\_INCOME\_TOTAL - Desvación Standar: 99639.77619155393  AMT\_INCOME\_TOTAL - Moda: 0 135000.0  dtype: float64  AMT\_INCOME\_TOTAL - Mediana: 157500.0 |
| --- |

* Es posible apreciar que la mayor cantidad de personas que se les ha aprobado la tarjeta de crédito no poseen un vehículo personal. Junto a esto es posible verificar que el nivel educacional de las personas que se les aprobó la tarjeta fue: "Secondary / secondary special". Además que estaban casadas y poseían una casa o departamento propio.
* Para el caso de la cantidad de hijos, la media y la desviación estándar no nos entregan una correcta información, esto se debe a que lógicamente no es posible tener un 0.5 hijos. Pero, si es posible comprobar que la mayor cantidad de personas que les otorgó una tarjeta de crédito no poseen hijos.
* Para el caso de la desviación estándar relacionado al sueldo, se puede ver si que si existe una cercanía de los datos a esta, pero también existe una tendencia a alejarse de esta. Indicando que los datos, si tiene una tendencia a estar dispersos

##### Resumen de variables con rechazo a poseer una tarjeta de crédito

| FLAG\_OWN\_CAR informacion:  count 4291  unique 2  top N  freq 2722  Name: FLAG\_OWN\_CAR, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  NAME\_EDUCATION\_TYPE informacion:  count 4291  unique 5  top Secondary / secondary special  freq 2890  Name: NAME\_EDUCATION\_TYPE, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  NAME\_FAMILY\_STATUS informacion:  count 4291  unique 5  top Married  freq 2914  Name: NAME\_FAMILY\_STATUS, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  NAME\_HOUSING\_TYPE informacion:  count 4291  unique 6  top House / apartment  freq 3782  Name: NAME\_HOUSING\_TYPE, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  FLAG\_EMAIL informacion:  count 4291  unique 2  top Y  freq 3840  Name: FLAG\_EMAIL, dtype: object  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  CNT\_CHILDREN - Media: 0.45374038685621065  CNT\_CHILDREN - Desvación Standar: 0.8037742360650749  CNT\_CHILDREN - Moda: 0 0  dtype: int64  CNT\_CHILDREN - Mediana: 0.0  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  AMT\_INCOME\_TOTAL - Media: 193430.4057329294  AMT\_INCOME\_TOTAL - Desvación Standar: 116439.22736656509  AMT\_INCOME\_TOTAL - Moda: 0 135000.0  dtype: float64  AMT\_INCOME\_TOTAL - Mediana: 171000.0 |
| --- |

* Es posible apreciar que el comportamiento para para todas las variables tienen el mismo comportamiento pero en cantidad de diferentes, tanto para personas que se les aprobó o no la tarjeta

### Identificando Outliers

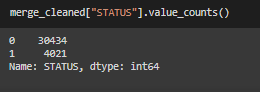
Para generar un dominio de datos ideal para trabajar se identificaron outliers y removidos:

| Con Outliers | Sin outliers |
| --- | --- |

## Operando sobre la información

A continuación se expondrán algoritmos que permitirá saber si es posible definir una clasificación

### Árbol de decisiones

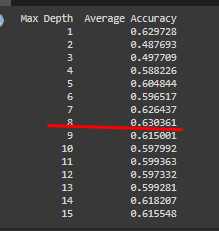


En base a esta información es posible verificar que existe una diferencia de 7.5 entre la cantidad de muestra entre el estatus 0 y el 1. Esto es necesario considerar porque se generará un peso de esta diferencia para la clase 1 en el árbol de decisión.

Se busco la mejor profundidad y en base al algoritmo

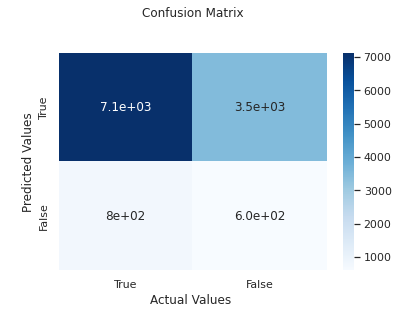


Dando como resultado:



Permitiendo tener el siguiente resultado: [link](https://drive.google.com/file/d/1pJjcuMT0H8oDeB5lQk22yyi1kHzAa_qo/view?usp=sharing)

#### Matriz de confusión



En base a los valores numéricos se concluye:

1. Posee una precisión bastante baja, por lo que es posible comprender que posee una gran dispersión de los datos. Y además es necesario comprobar que considerar casos positivos, el status 0 o 1.
2. El recall de la muestra indica que no es tan eficaz para identificar correctamente a las personas que se le otorgó el crédito.
3. Respecto al accuracy, ya contempla aumentar, y nos permite saber que está cercano a una estimación "positiva".
4. Al ser medianamente alta la especificidad nos permite entender que el algoritmo es factible para identificar los casos negativos.
5. Para el caso del f1-score, es posible decir que el modelo no logra clasificar correctamente.

**Para la medida adicional que se obtiene a través del reporte de clasificación de skelearn:**

1. Es posible verificar que el modelo permite verificar la etiqueta 0, que se le otorgará el crédito eficazmente.
2. El modelo tiene un déficit al momento de clasificar la etiqueta 1, que no se le otorgará el crédito.
   1. Esto se puede deber al soporte que posee, osea la cantidad de elementos que se usaron para construir el modelo.

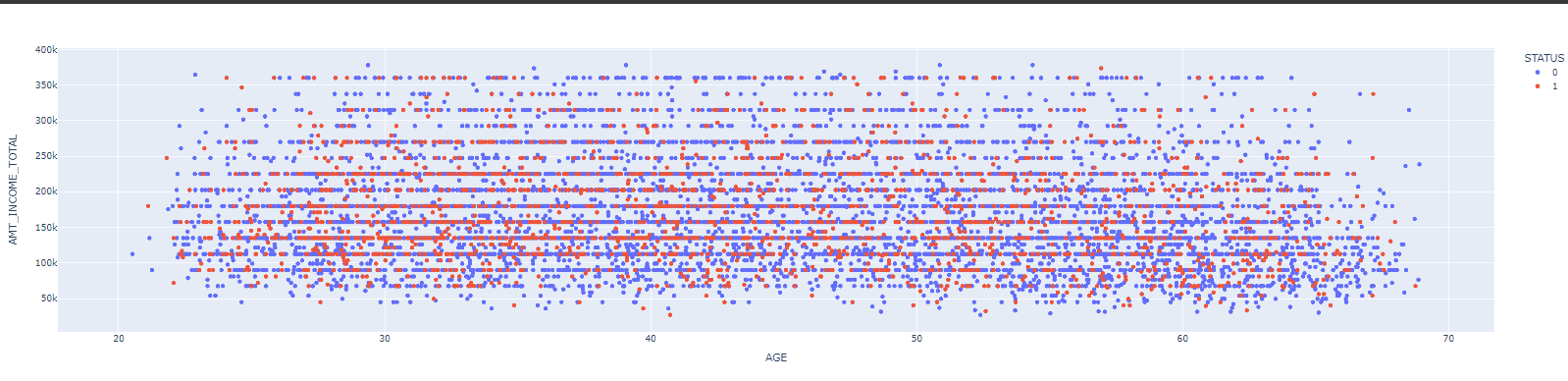
En base a la gráfica, se concluye::

1. Parte la segmentación del modelo identificando si posee o una propiedad. Continuando así con la diferenciación de los sueldos.
2. Así además las variables con mayor segmentación están siendo:
   1. El auto propio.
   2. La cantidad de hijos.
   3. Número de miembros en la familia.
   4. La ocupación.
   5. El género.
   6. El tipo de casa.

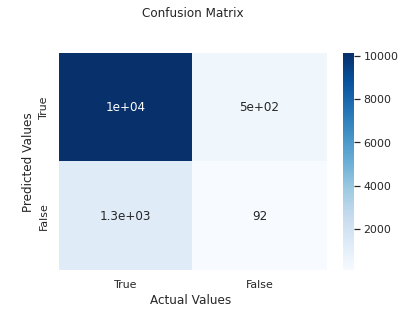
De esta forma es posible comprender que las variables que más predominan al momento de generar un correcto etiquetado son las señaladas.

### KNN

#### AMT\_INCOME\_TOTAL vs AGE



##### Matriz de confusión



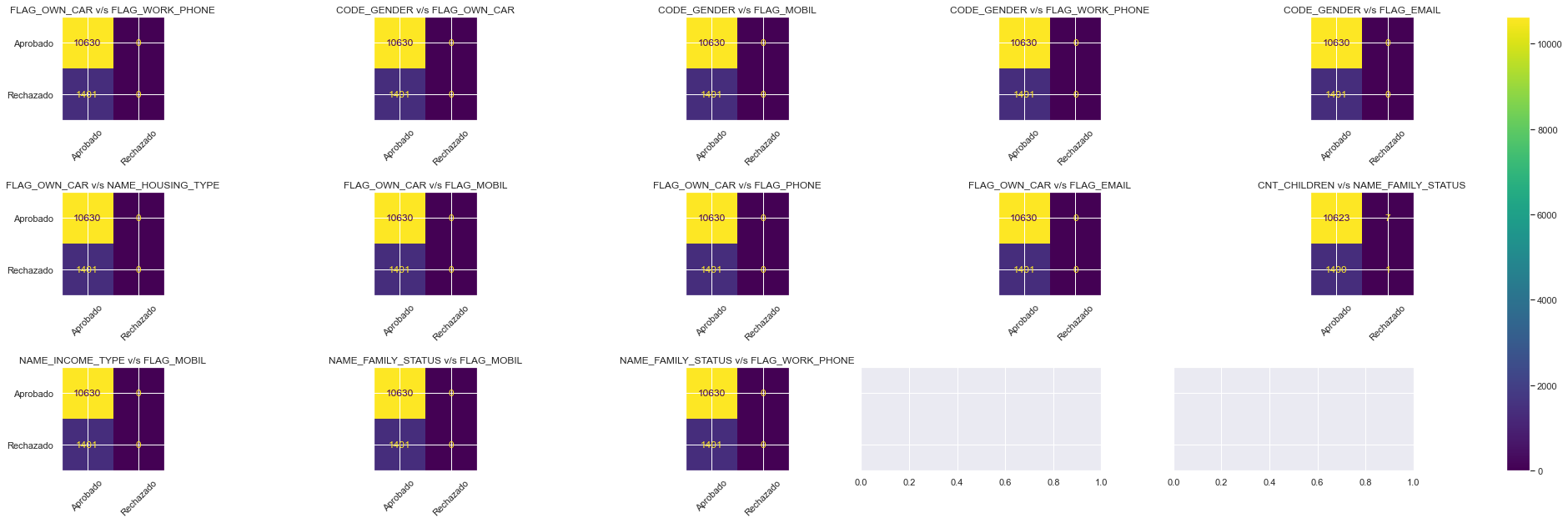
1. Se observa que el modelo presenta baja Precisión y bajo Recall con lo cual un bajo porcentaje de casos positivos fueron detectados. Esto debido a un alto nivel de dispersión en los datos. Dicho resultado también se traduce en el f1-score el cual resume la Precisión y Recall en una sola métrica.
2. La especificidad tiene un resultado considerablemente alto con lo cual el algoritmo ha clasificado correctamente los casos negativos.
3. La exactitud tiene un valor considerablemente alto con un la mayoría de los casos positivos fueron correctamente clasificados.

#### All's vs All's

A diferencia de la anterior que solo se compararon variables especfiicas, en este caso, se comparon todas las variables vs todas las variables menos por si mismo.

Y se obtuvieron los siguientes resultados:

##### Matriz de confusión



1. En base a lo que se ha podido revisar tanto con imágenes y valores numéricos, es posible apreciar que respecto a la etiqueta 0 (se le aprueba una tarjeta de crédito) teniendo una alta precisión y un alto recall, el modelo es capaz de trabajar correctamente con esta etiqueta
2. Para el caso contrario con la etiqueta 1 (no se aprueba una tarjeta de crédito) el modelo no la detecta del todo bien, pero, cuando lo hace, existe la seguridad que es correcta.

Notas:

* Cabe mencionar que existe un problema en la cantidad de datos, ya que la diferencia entre la cantidad de la prueba de la etiqueta 0 y 1, es casi de 9 veces.
* Se presentan varias comparaciones con la idea de estudiar el comportamiento de acuerdo con KNN, Donde se puede apreciar que con un KN = 2 es posible obtener los mismo resultados o mejor que a uno de 14