|  |
| --- |
| Modelo de originación de crédito para predecir el comportamiento de pago de nuevos créditos. |
| Bank Five International  Claudia Villegas  Barbara Herrera  Isabel Pilar  Miguel Peña  Alexis Tapia |
|  |



Contenido

[1. Introducción 4](#_Toc98096572)

[2. Objetivo 5](#_Toc98096573)

[2.1. Objetivo General 5](#_Toc98096574)

[2.2. Objetivos Específicos 5](#_Toc98096575)

[3. Fuente de información 6](#_Toc98096576)

[3.1. Atributos para análisis 6](#_Toc98096577)

[4. Población de estudio 7](#_Toc98096578)

[4.1. Créditos de Consumo 7](#_Toc98096579)

[4.2. Vector Objetivo 8](#_Toc98096580)

[5. Preprocesamiento de datos 9](#_Toc98096581)

[5.1. Análisis de nulos 9](#_Toc98096582)

[5.2. Tratamiento de nulos 11](#_Toc98096583)

[6. Análisis Univariado 12](#_Toc98096584)

[7. Correlaciones 15](#_Toc98096585)

[8. Recodificación y Transformación de variables 17](#_Toc98096586)

[9. Análisis bivariado 18](#_Toc98096587)

[10. Selección de variables 19](#_Toc98096588)

[11. Premodelamiento 21](#_Toc98096589)

[12. Modelamiento 22](#_Toc98096590)

[12.1. Modelos candidatos. 22](#_Toc98096591)

[12.2. Resultados 24](#_Toc98096592)

[12.3. Importancia de las variables 26](#_Toc98096593)

[13. Punto de Corte 27](#_Toc98096594)

[14. Implementación 30](#_Toc98096595)

[15. Conclusiones 31](#_Toc98096596)

[16. Anexos 32](#_Toc98096597)

[16.1. Herramientas de trabajo 32](#_Toc98096598)

[16.2. Detalle de atributos 32](#_Toc98096599)

[16.3. Detalle de nulos 37](#_Toc98096600)

[16.4. Gráficos de análisis univariado 39](#_Toc98096601)

[16.5. Categorizaciones 49](#_Toc98096602)

[16.6. Gráficos de análisis bivariado 55](#_Toc98096603)

[16.7. Marco teórico de algoritmos utilizados 59](#_Toc98096604)

[16.8. Reportes de resultados de los modelos 61](#_Toc98096605)

[16.9. Búsqueda hiperparámetros modelos seleccionados. 64](#_Toc98096606)

[16.9.1. Regresión Logística. 64](#_Toc98096607)

[16.9.2. Grandient boosting. 64](#_Toc98096608)

[16.10. Bibliografía 65](#_Toc98096609)

# Introducción

Un modelo de originación de crédito busca medir y discriminar entre los clientes que incumplen y no incumplen en los pagos mensuales de créditos, con la finalidad de identificar a los clientes mal pagadores, es así como se podrá mejorar la calidad de la cartera crediticia y tener una herramienta de apoyo a la gestión de evaluación de créditos.

Un modelo es una herramienta estadística basada en atributos del cliente y del crédito que permiten clasificar a los clientes según el resultado de la evaluación del crédito, asignando perfiles que logren identificar el riesgo de incumplimiento en el pago de un crédito.

Este documento presenta la metodología usada para la construcción del modelo, procesos previos y sustentos estadísticos. Además, se analizan cada uno de los resultados obtenidos para mejorar la comprensión y razonabilidad del modelo.

Finalmente, se concluye a partir de las métricas necesarias que permiten concluir que el modelo propuesto es capaz de reconocer características de diferentes atributos y asociarlas a un comportamiento de pago.

# Objetivo

## Objetivo General

* Construir un modelo predictivo para la empresa que permita estimar la probabilidad de que un cliente sea buen o mal pagador, para apoyar el proceso de evaluación de créditos.

## Objetivos Específicos

* Identificar variables que tenga mayor incidencia en la probabilidad de incumplimiento de pago de un crédito.
* Obtener un modelo que permitan estimar la probabilidad de incumplimiento de pago de los clientes, según características del cliente.
* Cuantificar beneficios y costos respecto a la implementación de un modelo.

# Fuente de información

## Atributos para análisis

La base de datos fue entregada por la Academia Desafío Latam para tiene un total de 121 atributos más la clase objetivo a modelar.

Los datos corresponden a otorgamientos de crédito de un banco internacional, en donde algunos de estos clientes registran morosidad, es decir no paga como se había comprometido. Por lo tanto, se requiere hacer un modelo para predecir el comportamiento de futuros buenos pagadores.

Los atributos consideran diferentes aspectos del cliente como son:

* Capacidad Económica
* Antecedentes Familiares
* Información crediticia
* Organización donde trabaja
* Patrimonio
* Situación Laboral
* Nivel Educacional
* Entre otros…

# Población de estudio

## Créditos de Consumo

Bank Five International otorga créditos de consumo (Cash loans) y créditos revolving (Revolving loans) a sus clientes como se observa en la ilustración.

Gráfico, Gráfico de barras

Descripción generada automáticamente

*Ilustración 1: Cartera de crédito*

En este contexto, el estudio se limitará al análisis de créditos de consumo, ya que, de la muestra, existen una mayoría de clientes que solicitan este tipo de producto en un 90% en comparación con un 10% de solicitudes de créditos revolvente.

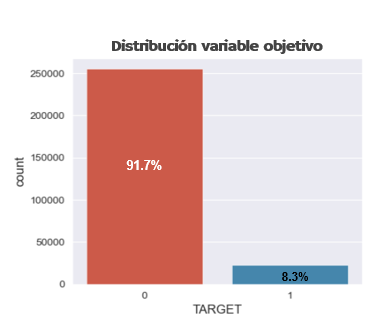
|  |  |  |
| --- | --- | --- |
| Tipo Crédito | Créditos | Porcentaje |
| Cash loans | 278.232 | 90.4% |
| Revolving loans | 29.279 | 9.5% |
| Total | 307.511 | 100% |

*Tabla 1: Frecuencia de créditos*

## Vector Objetivo

Se define el vector objetivo como una variable dicotómica que toma los siguientes valores:

De acuerdo a lo definido en el punto anterior, a base de datos contiene 278.232 registros de los cuales un 91,7% corresponden a clientes que son buenos pagadores, mientras que el 8,3% restante corresponden a clientes con dificultades para pagar.



*Ilustración 2: Variable dependiente a modelar*

# Preprocesamiento de datos

## Análisis de nulos

Los valores ausentes son muy importantes a la hora de crear modelos, porque estos pueden verse muy influenciados por observaciones faltantes o nulas, como se observa en la visualización de los datos.

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

*Ilustración 3: Visualización de nulos en los datos*

Como parte del preprocesado de datos, debe aprenderse con los datos de entrenamiento, su identificación (valores nulos) se realizó antes de separar los datos para asegurar las imputaciones posteriores, y dado que la data presenta un gran porcentaje de valores nulos identificados como **NaN** (**Ver Anexo 16.3:** Tabla con porcentaje de nulos de cada columna), se definió que las columnas que superaran un umbral de un 45% de nulos fueran eliminados.

|  |  |  |
| --- | --- | --- |
| **COMMONAREA\_AVG** | **COMMONAREA\_MODE** | **COMMONAREA\_MEDI** |
| **NONLIVINGAPARTMENTS\_AVG** | **NONLIVINGAPARTMENTS\_MODE** | **NONLIVINGAPARTMENTS\_MEDI** |
| **FONDKAPREMONT\_MODE** | **LIVINGAPARTMENTS\_MEDI** | **LIVINGAPARTMENTS\_AVG** |
| **LIVINGAPARTMENTS\_MODE** | **FLOORSMIN\_AVG** | **FLOORSMIN\_MODE** |
| **FLOORSMIN\_MEDI** | **YEARS\_BUILD\_AVG** | **YEARS\_BUILD\_MEDI** |
| **YEARS\_BUILD\_MODE** | **OWN\_CAR\_AGE** | **LANDAREA\_MEDI** |
| **LANDAREA\_AVG** | **LANDAREA\_MODE** | **BASEMENTAREA\_MEDI** |
| **BASEMENTAREA\_AVG** | **BASEMENTAREA\_MODE** | **EXT\_SOURCE\_1** |
| **NONLIVINGAREA\_AVG** | **NONLIVINGAREA\_MODE** | **NONLIVINGAREA\_MEDI** |
| **ELEVATORS\_MODE** | **ELEVATORS\_AVG** | **ELEVATORS\_MEDI** |
| **WALLSMATERIAL\_MODE** | **APARTMENTS\_AVG** | **APARTMENTS\_MEDI** |
| **APARTMENTS\_MODE** | **ENTRANCES\_AVG** | **ENTRANCES\_MEDI** |
| **ENTRANCES\_MODE** | **HOUSETYPE\_MODE** | **LIVINGAREA\_MEDI** |
| **LIVINGAREA\_MODE** | **LIVINGAREA\_AVG** | **FLOORSMAX\_MEDI** |
| **FLOORSMAX\_MODE** | **FLOORSMAX\_AVG** | **YEARS\_BEGINEXPLUATATION\_MEDI** |
| **YEARS\_BEGINEXPLUATATION\_MODE** | **YEARS\_BEGINEXPLUATATION\_AVG** | **TOTALAREA\_MODE** |
| **EMERGENCYSTATE\_MODE** |  |  |

*Tabla 2: Atributos con valores nulos que superan umbral 45%*

Por consiguiente, en la tabla descrita anteriormente, existen 49 atributos de un total de 121 variables que superan este umbral y fueron eliminados por considerarse variables que no aportan información relevante para el caso de estudio (características del lugar donde vive el cliente), además se desconoce el atributo original desde donde se obtuvieron las estadísticas ya que contienen los sufijos AVG (Promedio), MODE (Moda) y MEDI (Mediana).

Se identifico como **XAN** otro valor nulo presente en la variable **ORGANIZATION\_TYPE** con un total de 52008 valores nulos, equivalente a un 18% de sus y dado que no supera el umbral definido, esta variable no fue eliminada en este proceso.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Los atributos restantes con valores nulos se analizaron para proceder a imputarlos.

## Tratamiento de nulos

Es importante tener en cuenta, las consideraciones para imputar valores ausentes al resto de las variables con valores faltantes, ya que, si esta es errónea, se corre el riesgo de introducir valores en variables que puedan tener mucha influencia en el modelo, por tanto, se analizaron por tipo, empleando imputar los valores por la mediana para las variables de tipo numéricas o imputar por 0 y para las variables categóricas el valor NaN fue reemplazado por Unknown.

|  |  |
| --- | --- |
| **Variable** | **Tratamiento de nulos** |
| EXT\_SOURCE\_2 | Reemplazo por mediana |
| EXT\_SOURCE\_3 |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | Recodificación por cero |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT |
| AMT\_REQ\_CREDIT\_BUREAU\_MON |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR |
| AMT\_ANNUITY |
| DAYS\_LAST\_PHONE\_CHANGE |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE |
| OCCUPATION\_TYPE | Recodificación por "Unknown" |
| NAME\_TYPE\_SUITE |

*Tabla 3: Variables con tratamiento de nulos*

Las variables EXT\_SOURCE\_2 (mediana 0.565) y EXT\_OURCE\_3 (mediana 0.537) fueron imputadas por su valor de mediana, las variables con sufijo AMT\_REQ\_CREDIT\_BUREAU, AMT\_ANNUITY, DAYS\_LAST\_PHONE\_CHANGE, DEF\_60\_CNT\_SOCIAL\_CIRCLE, OBS\_60\_CNT\_SOCIAL\_CIRCLE, DEF\_30\_CNT\_SOCIAL\_CIRCLE y OBS\_60\_CNT\_SOCIAL\_CIRCLE fueron imputadas con el valor 0 por tener el mismo porcentaje de nulos.

Estas imputaciones realizadas a la base de entrenamiento se deben aplican posteriormente al conjunto de test.

# Análisis Univariado

El análisis con gráficos (Ver Anexo 16.4: Gráficos de análisis univariado) que permitió visualizar la distribución de los atributos. Las variables estudiadas (Ver Anexo 16.2: Detalle de los atributos) contenían información de la capacidad económica del cliente, antecedentes familiares, personales, comportamiento crediticio, situación laboral, organización donde trabaja, nivel educación y características del lugar donde vive el cliente, estas últimas características fueron eliminadas en su mayoría por contar con un porcentaje importante de valores nulos.

En la siguiente tabla se agrupa la información indicada anteriormente.

|  |  |  |
| --- | --- | --- |
| Información de la capacidad económica:   * AMT\_INCOME\_TOTAL * NAME\_INCOME\_TYPE | Información del préstamo:   * AMT\_CREDIT * AMT\_ANNUITY * AMT\_GOODS\_PRICE | Comportamiento crediticio:   * EXT\_SOURCE\_2 * EXT\_SOURCE\_3 |
| Situación laboral:   * DAYS\_EMPLOYED * OCCUPATION\_TYPE | **Organización donde trabaja:**   * ORGANIZATION\_TYPE | **Nivel educacional:**   * NAME\_EDUCATION\_TYPE |
| Información personal:   * REGION\_POPULATION\_RELATIVE * AGE * FLAG\_EMP\_PHONE * FLAG\_WORK\_PHONE * FLAG\_PHONE * DAYS\_LAST\_PHONE\_CHANGE * CODE\_GENDER * NAME\_TYPE\_SUITE * NAME\_FAMILY\_STATUS * NAME\_HOUSING\_TYPE * FLAG\_MOBIL | **Evaluación financiera:**   * DAYS\_REGISTRATION * DAYS\_ID\_PUBLISH * REGION\_RATING\_CLIENT * REG\_CITY\_NOT\_LIVE\_CITY * REG\_CITY\_NOT\_WORK\_CITY * LIVE\_CITY\_NOT\_WORK\_CITY * OBS\_30\_CNT\_SOCIAL\_CIRCLE * DEF\_30\_CNT\_SOCIAL\_CIRCLE * OBS\_60\_CNT\_SOCIAL\_CIRCLE * DEF\_60\_CNT\_SOCIAL\_CIRCLE * AMT\_REQ\_CREDIT\_BUREAU\_HOUR * AMT\_REQ\_CREDIT\_BUREAU\_DAY * AMT\_REQ\_CREDIT\_BUREAU\_WEEK * AMT\_REQ\_CREDIT\_BUREAU\_MON * AMT\_REQ\_CREDIT\_BUREAU\_QRT * AMT\_REQ\_CREDIT\_BUREAU\_YEAR * FLAG\_CONT\_MOBILE, FLAG\_EMAIL * REGION\_RATING\_CLIENT\_W\_CITY * WEEKDAY\_APPR\_PROCESS\_START * HOUR\_APPR\_PROCESS\_START * REG\_REGION\_NOT\_LIVE\_REGION * REG\_REGION\_NOT\_WORK\_REGION * LIVE\_REGION\_NOT\_WORK\_REGION * FLAG\_DOCUMENT\_2, FLAG\_DOCUMENT\_3, FLAG\_DOCUMENT\_4, FLAG\_DOCUMENT\_5, FLAG\_DOCUMENT\_6, FLAG\_DOCUMENT\_7, FLAG\_DOCUMENT\_8, FLAG\_DOCUMENT\_9, FLAG\_DOCUMENT\_10, FLAG\_DOCUMENT\_11, FLAG\_DOCUMENT\_12, FLAG\_DOCUMENT\_13, FLAG\_DOCUMENT\_14, FLAG\_DOCUMENT\_15, FLAG\_DOCUMENT\_16, FLAG\_DOCUMENT\_17, FLAG\_DOCUMENT\_18, FLAG\_DOCUMENT\_19, FLAG\_DOCUMENT\_20 y FLAG\_DOCUMENT\_21 | |
| Antecedentes familiares:   * CNT\_CHILDREN * CNT\_FAM\_MEMBERS |
| Patrimonio:   * FLAG\_OWN\_CAR * FLAG\_OWN\_REALTY |

*Tabla 4: Análisis de atributos*

Del análisis se presentan las principales características de las variables analizadas:

* Se pudo inferir que existían atributos que definían características que corresponden a categorías y se identificaban como variables del tipo int, por lo que se realizó la transformación del tipo de dato a tipo object para su mejor interpretación.
* Se transformaron atributos con valores negativos y reprocesadas como la edad, cantidad de días trabajados (DAYS\_EMPLOYED), hace cuantos días antes de la postulación cambia número de teléfono (DAYS\_LAST\_PHONE\_CHANGE), además de ser transformadas a categóricas en los dos últimos casos. De esta conversión la edad promedio de los clientes fue de 44 años.
* En variables como AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY, AMT\_GOODS\_PRICE fueron transformados a su valor en logaritmo para representar una mejor distribución de la variable dado que contaban con valores atípicos que sesgarían el análisis.
* Se eliminaron variables que no aportaban significancia al análisis como por ejemplo el día de la semana WEEKDAY\_APPR\_PROCESS\_START u hora en que el cliente solicito el préstamo o quien lo acompaña NAME\_TYPE\_SUITE cuando realiza esta solicitud porque prefiere no ir acompañado.
* También fueron eliminadas variables como OBS\_30\_CNT\_SOCIAL\_CIRCLE y OBS\_60\_CNT\_SOCIAL\_CIRCLE, que describían los mismo que DEF\_30\_CNT\_SOCIAL\_CIRCLE y DEF\_60\_CNT\_SOCIAL\_CIRCLE respecto de cuantas veces registro mora su entorno.
* Fueron eliminadas variables que estaban muy sesgadas a una sola categoría como por ejemplo FLAG\_MOBIL, FLAG\_CONT\_MOBILE, FLAG\_EMAIL, que indicaron que los clientes entregaron un número celular de contacto, se les pudo contactar a dicho número, pero no entregaron un email de contacto ni un teléfono de casa.
* Los clientes indicaron que la dirección permanente es la misma que la dirección de contacto y la misma para dirección de trabajo.
* La mayoría de los clientes fueron mujeres que solicitaron el préstamo.
* Muchos de los clientes no tienen automóvil, pero si son propietarios de una casa o departamento, además de indicar que son casados y que trabajan a la hora de solicitar el préstamo.
* El nivel educacional de la muestra contaba con estudios secundarios.
* Se recodificaron variables como por ejemplo NAME\_INCOME\_TYPE, ORGANIZATION\_TYPE, NAME\_EDUCATION\_TYPE, NAME\_FAMILY\_STATUS, NAME\_HOUSING\_TYPE, OCCUPATION\_TYPE para disminuir la cantidad de valores en las mismas analizando el porcentaje de malos por atributo.
* En el caso de las variables ORGANIZATION\_TYPE (con 58 categorías) y OCCUPATION\_TYPE (18 categorías) fueron recategorizadas a 8 y 5 respectivamente para disminuir la cantidad de valores dentro de las mismas. Para agruparlas se analizó la tasa de malos pagadores agrupando categorías que tenían el mismo porcentaje.
* En resumen, de este análisis, fueron eliminadas 16 atributos numéricos y 29 variables categóricas. Posteriormente, de la correlación entre atributos y análisis bivariado fueron eliminadas otras características que se explican en los siguientes apartados.

# Correlaciones

Se calcula la correlación de Pearson entre los atributos cuantitativos, La siguiente imagen muestra las variables que presentan una correlación mayor a 0,6 en valor absoluto.

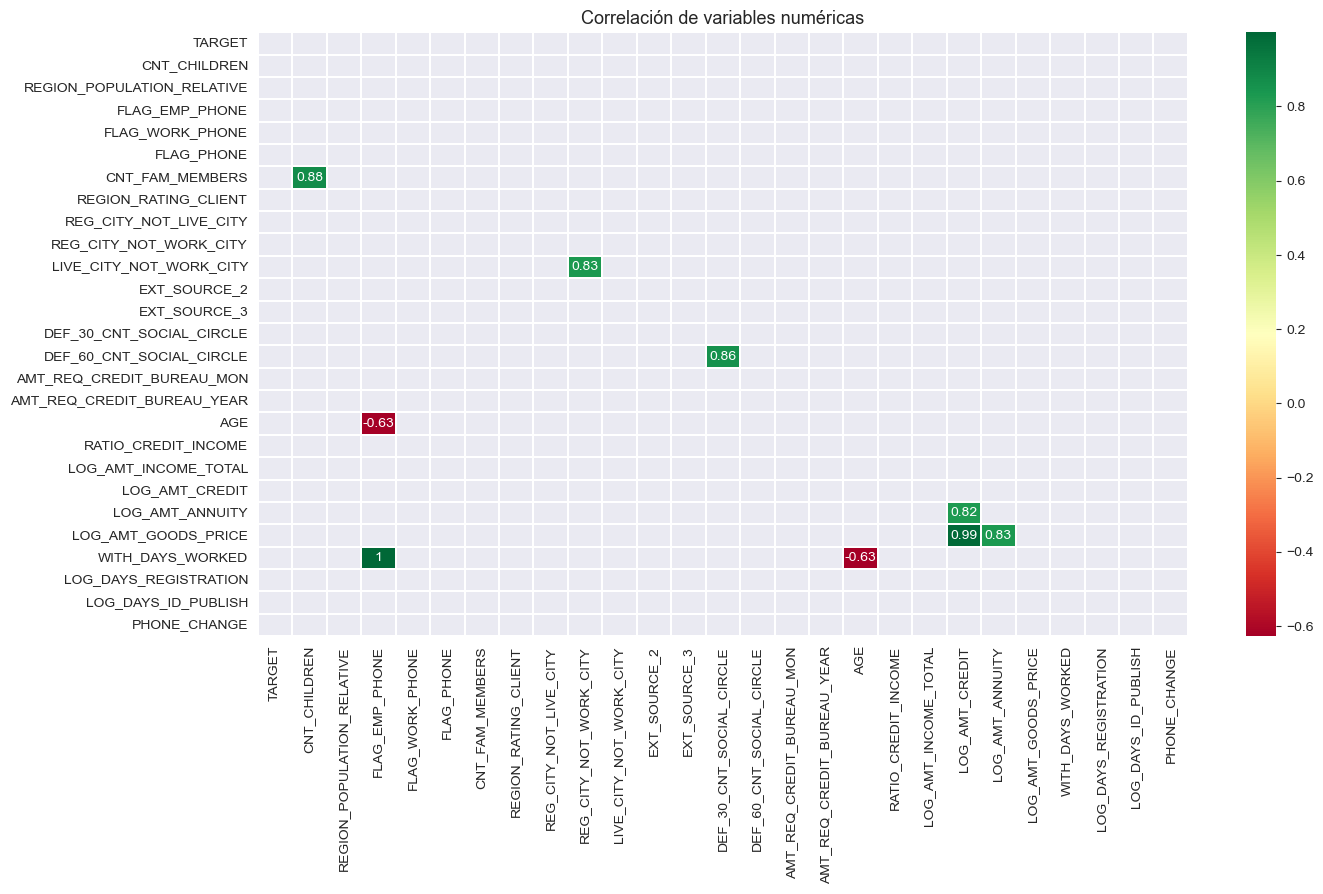
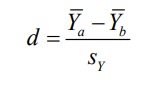
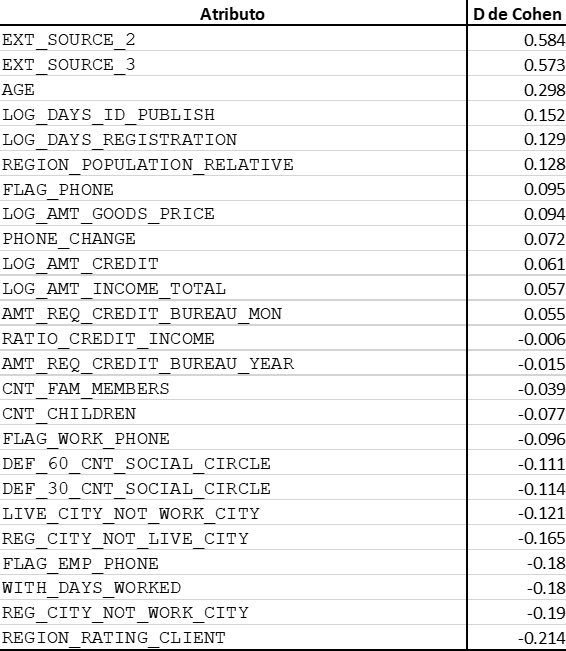


Ilustración 4: Matriz de correlación

Para calcular la asociación entre el target y los atributos cuantitativos utilizamos el índice de asociación d de Cohen que se calcula mediante la siguiente fórmula:

Los resultados se muestran en la tabla que se muestra a continuación:



*Tabla 5: Correlación con variable objetivo*

# Recodificación y Transformación de variables

Durante el análisis de datos se transformaron 10 variables, se creó 1 y se recodificaron 6.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **N°** | **Variables Transformadas** |  | **N°** | **Variables Creadas** |  | **N°** | **Variables Recodificadas** |
| **1** | **DAYS\_BIRTH** | **1** | **RATIO\_CREDIT\_INCOME** | **1** | **NAME\_INCOME\_TYPE** |
| **2** | **AMT\_INCOME\_TOTAL** |  | | **2** | **NAME\_EDUCATION\_TYPE** |
| **3** | **AMT\_CREDIT** | **3** | **NAME\_FAMILY\_STATUS** |
| **4** | **AMT\_ANNUITY** | **4** | **NAME\_HOUSING\_TYPE** |
| **5** | **RATIO\_CREDIT\_INCOME** | **5** | **OCCUPATION\_TYPE** |
| **6** | **AMT\_GOODS\_PRICE** | **6** | **ORGANIZATION\_TYPE** |
| **7** | **DAYS\_EMPLOYED** |  | |
| **8** | **DAYS\_REGISTRATION** |
| **9** | **DAYS\_ID\_PUBLISH** |
| **10** | **DAYS\_LAST\_PHONE\_CHANGE** |

*Tabla 6: Recodificación y transformación de variables*

Para mayor detalle del análisis Ver Anexo 16.5

# Análisis bivariado

Para las variables cuantitativas, se gráfica la distribución de la variable por cada categoría del target (buen pagador o mal pagador)como se puede ver en los gráficos del Anexo 16.6.

Se puede apreciar que la variable edad tiene una distribución distinta entre las categorías del target, mientras el resto de las variables, la distribución éstas son similares en cada una de las categorías del target.

Para las variables categóricas, en primero lugar se ha calculado el **porcentaje de malos pagadores** por cada categoría de la variable para luego comparar los porcentajes calculados. Los gráficos se muestran en el Anexo 16.6.

Se puede apreciar que para la variable FLAG\_OWN\_REALTY tiene un porcentaje similar de malos pagadores en cada una de sus categorías. Por otro lado, el resto de las variables tiene un porcentaje de malos pagadores diferente en cada una de las categorías.

# Selección de variables

A partir de los análisis realizados, se llega a un conjunto de 29 variables, las cuales se consideran para entrar a la etapa de modelamiento.

* Cabe destacar que dentro de los criterios utilizados para la selección de variables se encuentran:
* **Missings**: Se eliminan variables con más del 45% de valores nulos.
* **Concentración**: Se descartan variables que se encuentran concentradas principalmente en un solo valor (por ejemplo, varios casos de variables asociadas a FLAG).
* **Cualitativo**: o criterio de negocios, variables que no es de interés que sean predictoras (por ejemplo, día de la semana).
* **Relación con variable objetivo**: De análisis bivariado se seleccionan variables que se aprecian se relacionan con el vector objetivo (o se descartan si no se observa variabilidad en relación a la variable objetivo).
* **Correlación entre variables**: Si hay variables con muy alta correlación se selecciona una de estas (este criterio no se aplica de forma tan estricta ya que modelos de ensambles no requieren que se cumpla esta condición)

Todos estos criterios son los que fueron abordados en las etapas presentadas en las secciones anteriores.

A continuación, se presentan las variables seleccionadas para modelar:

|  |  |  |
| --- | --- | --- |
| N | Variable | Descripción |
| 1 | **FLAG\_OWN\_CAR** | Binario tiene auto |
| 2 | **FLAG\_PHONE** | Binario tiene fono de contacto |
| 3 | **FLAG\_WORK\_PHONE** | Binario tiene fono hogar |
| 4 | **PHONE\_CHANGE** | Cambia número de teléfono |
| 5 | **AMT\_REQ\_CREDIT\_BUREAU\_YEAR** | Cantidad de consultas bureau año |
| 6 | **AMT\_REQ\_CREDIT\_BUREAU\_MON** | Cantidad de consultas bureau mes |
| 7 | **CNT\_CHILDREN** | Cantidad de hijos |
| 8 | **CNT\_FAM\_MEMBERS** | Cantidad de miembros familiares |
| 9 | **NAME\_HOUSING\_TYPE** | Condición habitacional |
| 10 | **LOG\_DAYS\_ID\_PUBLISH** | Días previos a la modificación de ID (logaritmo). |
| 11 | **LOG\_DAYS\_REGISTRATION** | Días previos a la modificación de antecedentes (logaritmo) |
| 12 | **AGE** | Edad |
| 13 | **REGION\_RATING\_CLIENT** | Evaluación región del cliente |
| 14 | **LOG\_AMT\_INCOME\_TOTAL** | Ingreso del cliente (logaritmo), |
| 15 | **REG\_CITY\_NOT\_WORK\_CITY** | Marca de asociación hogar trabajo |
| 16 | **LIVE\_CITY\_NOT\_WORK\_CITY** | Marca domicilio laboral no concuerda |
| 17 | **REG\_CITY\_NOT\_LIVE\_CITY** | Marca domicilio no concuerda |
| 18 | **WITH\_DAYS\_WORKED** | Tiene días trabajados |
| 19 | **LOG\_AMT\_CREDIT** | Monto préstamo (logaritmo) |
| 20 | **NAME\_EDUCATION\_TYPE** | Nivel educacional |
| 21 | **EXT\_SOURCE\_2** | Puntaje fuente externa 2 |
| 22 | **EXT\_SOURCE\_3** | Puntaje fuente externa 3 |
| 23 | **CODE\_GENDER** | Sexo |
| 24 | **NAME\_FAMILY\_STATUS** | Situación familiar |
| 25 | **ORGANIZATION\_TYPE** | Tipo empresa laboral |
| 26 | **NAME\_INCOME\_TYPE** | Tipo de Ingreso |
| 27 | **OCCUPATION\_TYPE** | Tipo de ocupación |
| 28 | **DEF\_30\_CNT\_SOCIAL\_CIRCLE** | Veces con más de 30 días de mora |
| 29 | **DEF\_60\_CNT\_SOCIAL\_CIRCLE** | Veces con más de 60 días de mora |

*Tabla 7: Variables seleccionadas modelo*

Una vez binarizadas las variables categóricas, se obtienen en total 44 variables que se utilizan para el modelamiento.

# Premodelamiento

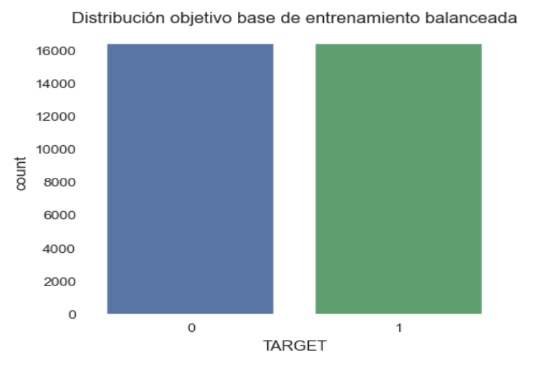
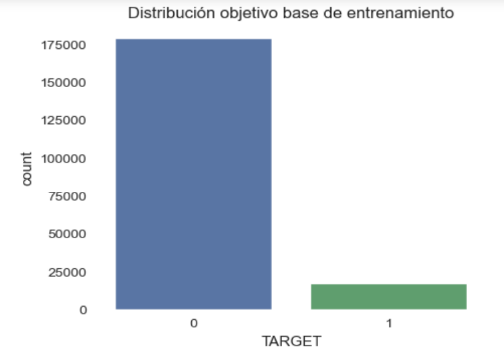
se divide la base en entrenamiento y validación (70% y 30%) con el uso del método *train\_test\_split.* Una vez generadas la base de entrenamiento se obtiene una base con 194.762 registros, correspondiente al 70% de la data original, la cual mantiene la distribución de target de la data original

Previo a la realización de los modelos se realiza una estandarización de los datos, para lo cual se utiliza el método *StandardScaler*.

Luego se realiza un balanceo de los datos de entrenamiento, debido a que la clase objetivo se encuentra desbalanceada, donde un 8.3% de los registros tienen marca de TARGET = 1 (correspondiente a clientes malos pagadores).

Para balancear la clase se consideran distintas metodologías, optando finalmente por realizar un submuestreo, es decir, seleccionar aleatoriamente registros de la clase 0, para que la base se encuentre mejor distribuida, para lo cual se utiliza el método *RandomUnderSampler* de sklearn.

Una vez que la base está balanceada, con 32.578 registros, con la distribución de target balanceada (50% clase 0 y 50% clase 1).



*Ilustración 5: Balaceo de conjunto de entrenamiento*

# Modelamiento

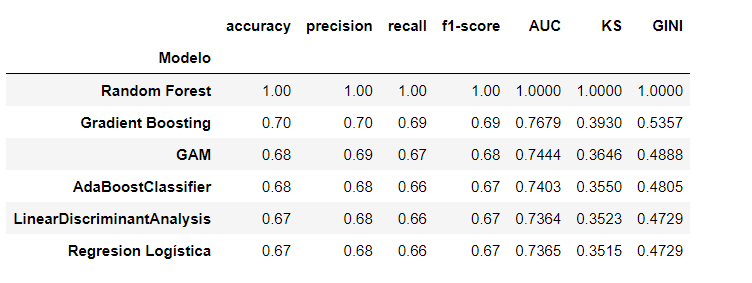
Dado el vector objetivo, se consideran distintos tipos de modelos de clasificación. Como primera alternativa se realiza un modelo de regresión logística, por ser un modelo que se utiliza en la industria crediticia al tener una mayor facilidad de interpretación. Además, se realizan pruebas con otras metodologías para evaluar si existen otros modelos con mejor capacidad de predicción. Dentro de los modelos a considerar se encuentran: Modelos aditivos generalizados, RandomForest, GradientBoosting, AdaBoosting y Análisis linear discriminante.

Dentro de los indicadores a utilizar para comparar los modelos se encuentran: precision, recall, f1-score, accuracy (generados a través de un reporte de clasificación), AUC (curva ROC), y adicionalmente los indicadores KS y GINI los cuales son indicadores que se utilizan en la industria para medir capacidad predictiva de modelos de créditos.

## Modelos candidatos.

A continuación, se presentan los resultados de los distintos modelos ejecutados, los cuales se comparan entre la base de entrenamiento y validación.

**Indicadores data de entrenamiento.**



*Tabla 8: Métricas de desempeño en datos de entrenamiento*

**Indicadores data de validación.**



*Tabla 9: Métricas de desempeño en datos de validación*

De los resultados anteriores se puede observar que los distintos modelos tienen indicadores similares con un buen poder predictivo, considerando la siguiente tabla como referencia:

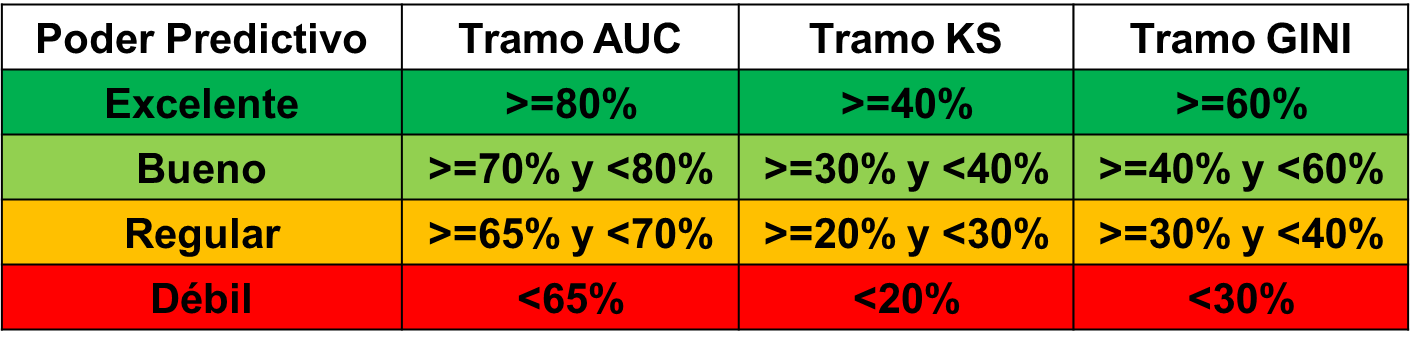


Tabla 10: Referencia Indicadores AUC, KS, GINI

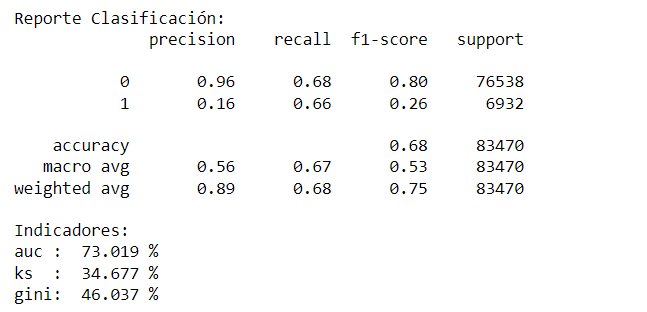
Como se puede apreciar de los resultados. Los modelos de gradient boosting, GAM, Regresión Logística y análisis lineal discriminante son los que tienen mejores indicadores, considerando que tienen un KS aproximado entre 35% y 36% al evaluarlos en la base de test.

A partir de esto se seleccionan como modelos candidatos el de regresión logística -a pesar de que modelo GAM un indicador KS levemente mejor y gradient boosting. Esto porque por una parte la regresión logística presenta una mejor opción considerando que es un modelo de más rápida ejecución e interpretabilidad y gradient boosting por ser el modelo con mejores indicadores. Con estos modelos se realiza la búsqueda de grilla para la búsqueda de hiperparametros óptimos, obteniendo como mejor modelo el de gradient boosting (Ver anexo: Búsqueda hiperparámetros modelos seleccionados).

## Resultados

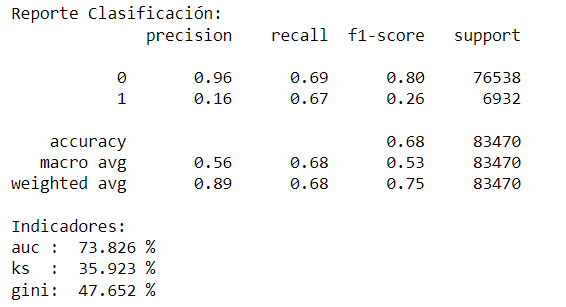
A continuación, se presentan el detalle de los resultados de los modelos candidatos:

**Regresión Logística:**

****

*Ilustración 6: Métricas de Regresión Logística*

**GradientBoosting:**

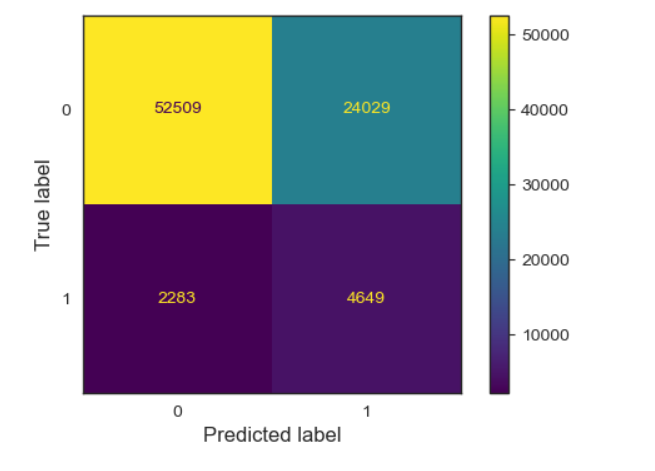


*Ilustración 7: Métricas de Gradient Boosting*

A partir de estos resultados se selecciona el modelo de gradient boosting.

Cabe destacar que, a pesar de los hiperparámetros, el indicador de precisión se mantiene bajo (lo cual se observó en todos los modelos como se mostró en la sección anterior), a pesar de esto, se consideran buenos modelos ya que su recall es mejor (sobre 0.5). En este sentido, la decisión de seleccionar este modelo pasa por el hecho de que de los casos que son malos pagadores, un 67% efectivamente lo es, es decir, aunque existen casos que son buenos pagadores y que el modelo considera que no lo son, se privilegia el poder aumentar la predicción de los malos pagadores (que pueden significar mayor pérdida para el negocio por créditos entregados que no serán pagados).

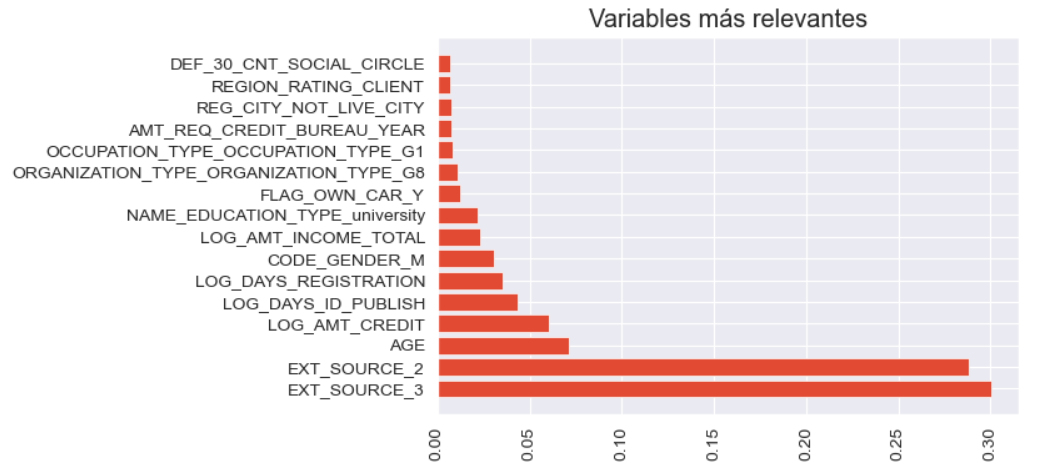
Lo anterior se puede apreciar más claramente en la matriz de confusión, donde de los 6.932 casos de malos pagadores, el modelo predice como mal pagador 4.649, pero también predice como mal pagador 24.029 casos que en realidad son buenos pagadores.



*Ilustración 8: Matriz de confusión*

## Importancia de las variables

Una vez seleccionado el modelo, se analiza la importancia de las variables obtenidas. Dada la cantidad de variables utilizadas en el modelo, se presentan las 15 más importantes:



*Ilustración 9: Importancia variables modelo seleccionado*

Es importante destacar que las variables que presentan más importancia en la predicción son EXT\_SOURCE\_3 y EXT\_SOURCE\_2, los cuales corresponden a un puntaje de externo del cliente. Del análisis bivariado se pudo observar que mayores puntajes están asociados a mejores pagadores.

Otras variables relevantes son la edad (AGE), donde del análisis bivariado se puede interpretar que a mayor edad es mejor pagador el cliente y el monto del crédito (ingresado como el logaritmo de este monto, LOG\_AMT\_CREDIT), donde también se puede inferir de análisis bivariado que mejores pagadores se concentran en casos donde se les otorgó más créditos.

Otras de las variables que se encuentran dentro de las 10 más relevantes son los días previos a la modificación del documento de identificación (LOG\_DAYS\_ID\_PUBLISH), cantidad de días previos a la modificación del último registro del cliente (LOG\_DAYS\_REGISTRATION), el sexo del cliente masculino (CODE\_GENDER\_M), el monto del ingreso del cliente (como logaritmo en variable LOG\_AMT\_CREDIT), el tipo de educación universitaria (NAME\_EDUCACTION\_TYPE\_university) y si posee auto (FLAG\_OWN\_CAR\_Y).

# Punto de Corte

Como se mencionó en la sección de resultados, a pesar de los buenos indicadores de predicción, existe un margen de error donde buenos pagadores se clasifican como malos pagadores. Para poder tener mayor flexibilidad, y así poder cubrir algún porcentaje de estos casos, se analiza los resultados en función de la probabilidad.

Para esto, se dividen los casos de la base de test en 20 tramos (cuantiles), en función de la probabilidad de ser mal pagador, y en función del número de casos de buenos y malos pagadores (reales) contenidos en cada tramo, lo cual permite apreciar cómo se discrimina entre buenos y malos pagadores a medida que aumenta la probabilidad.

Además, se realiza una simulación de los beneficios asociados a entregar un crédito a un buen pagador, y el costo asociado a ser un mal pagador, considerando lo siguiente:

* 15% de beneficio para créditos buenos.
* 135% de costos para créditos malos (considera por ejemplo además del crédito otros costos asociados como los de cobranza).

En base a esto y utilizando el monto promedio de los créditos de consumo de buenos y malos pagadores se estima:

* Ganancia de un crédito bueno: 94.869
* Pérdida de un crédito malo: 781.109

En base al número de buenos y malos pagadores en cada tramo, la ganancia y perdida estimada, es posible obtener una utilidad promedio para cada tramo como se puede apreciar en la siguiente tabla (donde tramo\_prob corresponde al tramo de probabilidad de ser mal pagador):



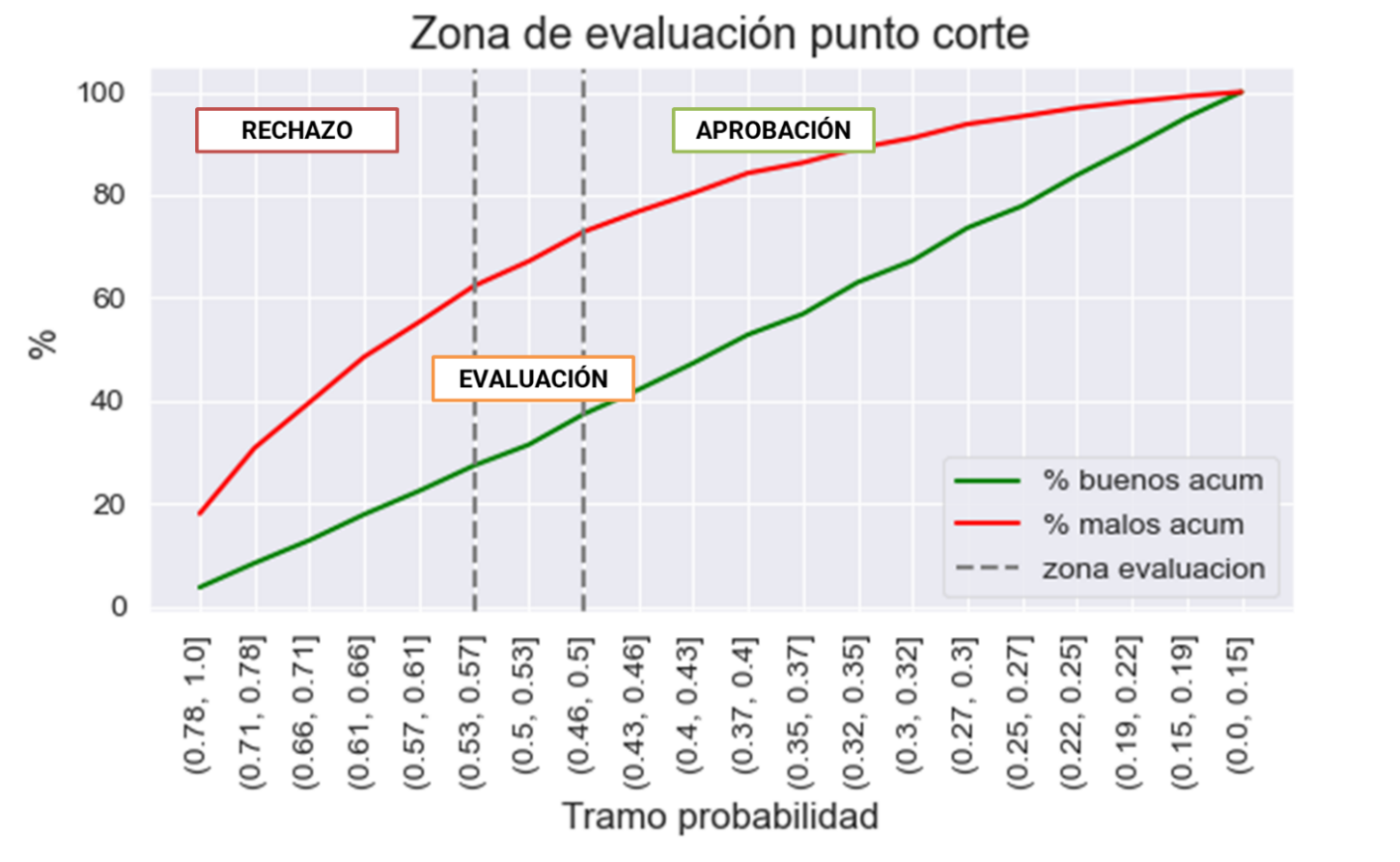
Tabla 11: Beneficio y Pérdida de otorgar crédito según probabilidad de comportamiento de pago.

De la tabla anterior se puede apreciar que las probabilidades menores a 0,53, se asocian a una utilidad esperada positiva (a menor tramo de probabilidad de ser mal pagador, mayor utilidad esperada), mientras que, a su vez, con probabilidades mayores a 0,53 se esperan utilidades negativas. Cabe recordar que el criterio de corte por defecto del modelo es 0,5, es decir, con una probabilidad sobre 0,5 el modelo clasificará a un cliente como mal pagador.

En consecuencia, es posible apreciar que existen intervalos que, si bien a priori clasificarían a un cliente como mal pagador, es posible re considerar esta clasificación dado que en el tramo la pérdida no es significativa. Así se observa que en el tramo (0,5 – 0,53] se espera una utilidad positiva baja, mientras que en el tramo (0,53 y 0,57] se espera una utilidad negativa baja. Por lo que se podría considerar en estos tramos no considerar inmediatamente a un cliente como mal pagador, si no que tener la opción de que se pueda evaluar de forma más detallada, para que con más antecedentes se pueda reevaluar el resultado del modelo (y poder considerar si entregar el crédito).

Por lo tanto, se propone definir una zona intermedia de evaluación manual por parte de algún ejecutivo, para aquellos casos que presenten una probabilidad entre (0,5 y 0,57], para considerar la aprobación del crédito.

En la siguiente figura es posible apreciar como el modelo discrimina entre buenos y malos pagadores, y cuál es la zona de evaluación intermedia propuesta para la clasificación de un cliente para la entrega de un crédito.



*Ilustración 10: Evaluación punto de corte*

# Implementación

Una vez definido el modelo, se importa el set de datos disponible para aplicarlo, junto al modelo serializado. Luego se realizan los ajustes necesarios para disponer de las variables necesarias (29 variables originales, 43 variables considerando binarizaciones) que permitan aplicar el modelo generado en estos datos. Cabe destacar que esta base no contiene la marca TARGET, por lo que es posible determinar el resultado del modelo, pero no validarlo.

La data de implementación cuenta con 48.744, de los cuales, 48.305 corresponden a créditos de consumo para los cuales se implementa el modelo.

Al realizar esto, se obtienen 32.584 (67.45%) donde el crédito es aprobado, 11.356 (23.51%) donde es rechazado y 4.365 casos donde se sugiere evaluar de forma exhaustiva si efectivamente el crédito debe ser rechazado.



*Ilustración 11: Resultados aplicación modelo*

# Conclusiones

A partir del presente desarrollo se ha podido constatar que si es posible predecir el comportamiento de pago de los clientes por medio de la implementación de un modelo de machine learning el cual permite identificar patrones desde un conjunto de atributos que permiten concluir de si un cliente será un buen o mal pagador a futuro.

Este modelo se debe incorporar como una herramienta adicional al flujo de proceso de evaluación de crédito, siendo un importante discriminador en la etapa de selección de clientes. Por lo tanto, la empresa deberá desplegar esfuerzos manuales solo para realizar evaluaciones manuales de algunos clientes en donde el modelo no haga una mayor diferenciación de posible comportamiento.

Este modelo deberá ser monitoreado mensualmente, para garantizar su efectividad en las evaluaciones y en caso de que se detecten variaciones en métricas de desempeño por debajo del umbral “bueno”, se deberá realizar una evaluación del modelo y/o posterior calibración de este.

# Anexos

## Herramientas de trabajo

* Jupyter Notebook
* Librerías: Pandas, Numpy, Sickit-learn

## Detalle de atributos

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Variable | | | Detalle | |
| 1 | **SK\_ID\_CURR** | | | Id préstamo realizado. | |
| 2 | **TARGET** | | | Variable objetivo, indica si el cliente presenta dificultades para pagar o no presenta dificultades. | |
| 3 | **NAME\_CONTRACT\_TYPE** | | | Identificación si es que el préstamo es en efectivo (cash) o en crédito revolvente (revolving). | |
| 4 | **CODE\_GENDER** | | | Sexo del cliente. | |
| 5 | **FLAG\_OWN\_CAR** | | | Indicador binario sobre la tenencia de automóvil por parte del cliente. | |
| 6 | **FLAG\_OWN\_REALTY** | | | Indicador binario sobre la propiedad de una casa o departamento por parte del cliente. | |
| 7 | **CNT\_CHILDREN** | | | Cantidad de hijos por parte del cliente. | |
| 8 | **AMT\_INCOME\_TOTAL** | | | Ingreso total del cliente. | |
| 9 | **AMT\_CREDIT** | | | Cantidad total del préstamo realizado | |
| 10 | **AMT\_ANNUITY** | | | Anualidad del préstamo. | |
| 11 | **AMT\_GOODS\_PRICE** | | | Para préstamos de consumo representa el precio de los bienes que se comprará con el préstamo. | |
| 12 | **NAME\_TYPE\_SUITE** | | | Quien acompaña al cliente cuando fue a solicitar el préstamo | |
| 13 | **NAME\_INCOME\_TYPE** | | | Tipo de ingreso por parte del cliente (empresario, asalariado, postnatal, entre otros). | |
| 14 | **NAME\_EDUCATION\_TYPE** | | | Máximo nivel educacional por parte del cliente. | |
| 15 | **NAME\_FAMILY\_STATUS** | | | Situación familiar del cliente. | |
| 16 | **NAME\_HOUSING\_TYPE** | | | Cuál es la situación habitacional del cliente. (arrienda, allegado, entre otros). | |
| 17 | **REGION\_POPULATION\_RELATIVE** | | | Población donde vive el cliente (la variable está normalizada, donde valores más altos significan que el cliente vive en una región más poblada). | |
| 18 | **DAYS\_BIRTH** | | | Edad del cliente cuando solicitó el préstamo. | |
| 19 | **DAYS\_EMPLOYED** | | | Cantidad de días trabajados previos a la postulación | |
| 20 | **DAYS\_REGISTRATION** | | | Cantidad de días previos a la última modificación de los registros del cliente previos a la postulación. | |
| 21 | **DAYS\_ID\_PUBLISH** | | | Cantidad de días previos a la modificación de su documento de identificación con el cual postulación al préstamo. | |
| 22 | **OWN\_CAR\_AGE** | | | Edad del auto del cliente. | |
| 23 | **FLAG\_MOBIL** | | | Da un teléfono celular de contacto:   * 1: YES * 0: NO | |
| 24 | **FLAG\_EMP\_PHONE** | | | Da un teléfono de trabajo de contacto:   * 1: YES * 0: NO | |
| 25 | **FLAG\_WORK\_PHONE** | | | Da un teléfono de hogar de contacto:   * 1: YES * 0: NO | |
| 26 | **FLAG\_CONT\_MOBILE** | | Se pudo contactar al cliente mediante el celular:   * 1: YES   0: NO | | |
| 27 | **FLAG\_PHONE** | | Da un teléfono contacto el cliente:   * 1: YES * 0: NO | | |
| 28 | **FLAG\_EMAIL** | | Da un email de contacto el cliente:   * 1: YES * 0: NO | | |
| 29 | **OCCUPATION\_TYPE** | | Cuál es la profesión del cliente. | | |
| 30 | **CNT\_FAM\_MEMBERS** | | Cuántos miembros familiares tiene el cliente. | | |
| 31 | **REGION\_RATING\_CLIENT** | | Evaluación interna (de Home Credit Group) sobre la región donde vive el cliente. | | |
| 32 | **REGION\_RATING\_CLIENT\_W\_CITY** | | Evaluación interna (de Home Crédito Group) sobre la región donde vive el cliente considerando ciudad. | | |
| 33 | **WEEKDAY\_APPR\_PROCESS\_START** | | Día hábil en el cual el cliente pide el préstamo. | | |
| 34 | **HOUR\_APPR\_PROCESS\_START** | | Hora aproximada de la solicitud de préstamo por parte del cliente. | | |
| 35 | **REG\_REGION\_NOT\_LIVE\_REGION** | | Identificador booleano si es que la dirección permanente del cliente no concuerda con la dirección de contacto:   * 1: different. * 0: same, at region level. | | |
| 36 | **REG\_REGION\_NOT\_WORK\_REGION** | | Identificador booleano si es que la dirección permanente del cliente no concuerda con la dirección de trabajo:   * 1: different. * 0: same, at region level). | | |
| 37 | **LIVE\_REGION\_NOT\_WORK\_REGION** | | Identificador booleano si es que la dirección de contacto del cliente no concuerda con la dirección del trabajo:   * 1: different. * 0: same, at region level). | | |
| 38 | **REG\_CITY\_NOT\_LIVE\_CITY** | | Identificador booleano si es que la dirección permanente no concuerda con la dirección de contacto:   * 1: different. * 0: same, at city level. | | |
| 39 | **REG\_CITY\_NOT\_WORK\_CITY** | | Identificador booleano si es que la dirección permanente no concuerda con la dirección del trabajo:   * 1: different. * 0: same, at city level. | | |
| 40 | **LIVE\_CITY\_NOT\_WORK\_CITY** | | Identificador booleano si es que la dirección de contacto del cliente no concuerda con la dirección del trabajo:   * 1: different. * 0: same, at city level. | | |
| 41 | **ORGANIZATION\_TYPE** | | Tipo de organización donde trabaja el cliente. | | |
| 42 | **EXT\_SOURCE\_1** | | Puntaje normalizado de fuente externa. | | |
| 43 | **EXT\_SOURCE\_2** | | Puntaje normalizado de fuente externa. | | |
| 44 | **EXT\_SOURCE\_3** | | | | Puntaje normalizado de fuente externa. |
| De esta tabla, desde la columna 45 a la , los nombres de los campos tienen los siguientes sufijos:   * AVG: Promedio. * MODE: Moda. * MEDI: Mediana.   Y la información corresponde a información normalizada sobre las características del lugar donde vive el cliente:   * Cantidad de habitaciones. * Tamaño del inmueble. * Áreas comunes. * Áreas de living. * Antigüedad en años de la construcción del edificio o vivienda. * Cantidad de ascensores. * Cantidad de entradas. * Estado del edificio. * Número del piso donde reside. | | | | | |
| 45 | **APARTMENTS\_AVG** | | | | Cantidad de habitaciones promedio. |
| 46 | **BASEMENTAREA\_AVG** | | | | Tamaño del inmueble promedio. |
| 47 | **YEARS\_BEGINEXPLUATATION\_AVG** | | | | Estado del edificio promedio. |
| 48 | **YEARS\_BUILD\_AVG** | | | | Antigüedad en años de la construcción del edificio o vivienda promedio. |
| 49 | **COMMONAREA\_AVG** | | | | Áreas comunes promedio. |
| 50 | **ELEVATORS\_AVG** | | | | Cantidad de ascensores promedio |
| 51 | **ENTRANCES\_AVG** | | | | Cantidad de entradas promedio. |
| 52 | **FLOORSMAX\_AVG** | | | | Número de pisos máximo promedio |
| 53 | **FLOORSMIN\_AVG** | | | | Número de pisos mínimo promedio |
| 54 | **LANDAREA\_AVG** | | | | Área del terreno promedio. |
| 55 | **LIVINGAPARTMENTS\_AVG** | | | | Áreas de living habitable promedio |
| 56 | **LIVINGAREA\_AVG** | | | | Áreas de terreno disponible promedio |
| 57 | **NONLIVINGAPARTMENTS\_AVG** | | | | Cantidad de habitaciones no habitable promedio. |
| 58 | **NONLIVINGAREA\_AVG** | | | | Área de terreno no disponible |
| 59 | **APARTMENTS\_MODE** | | | |  |
| 60 | **BASEMENTAREA\_MODE** | | | |  |
| 61 | **YEARS\_BEGINEXPLUATATION\_MODE** | | | |  |
| 62 | **YEARS\_BUILD\_MODE** | | | |  |
| 63 | | **COMMONAREA\_MODE** | | |  |
| 64 | | **ELEVATORS\_MODE** | | |  |
| 65 | | **ENTRANCES\_MODE** | | |  |
| 66 | | **FLOORSMAX\_MODE** | | |  |
| 67 | | **FLOORSMIN\_MODE** | | |  |
| 68 | | **LANDAREA\_MODE** | | |  |
| 69 | | **LIVINGAPARTMENTS\_MODE** | | |  |
| 70 | | **LIVINGAREA\_MODE** | | |  |
| 71 | | **NONLIVINGAPARTMENTS\_MODE** | | |  |
| 72 | | **NONLIVINGAREA\_MODE** | | |  |
| 73 | | **APARTMENTS\_MEDI** | | |  |
| 74 | | **BASEMENTAREA\_MEDI** | | |  |
| 75 | | **YEARS\_BEGINEXPLUATATION\_MEDI** | | |  |
| 76 | | **YEARS\_BUILD\_MEDI** | | |  |
| 77 | | **COMMONAREA\_MEDI** | | |  |
| 78 | | **ELEVATORS\_MEDI** | | |  |
| 79 | | **ENTRANCES\_MEDI** | | |  |
| 80 | | **FLOORSMAX\_MEDI** | | |  |
| 81 | | **FLOORSMIN\_MEDI** | | |  |
| 82 | | **LANDAREA\_MEDI** | | |  |
| 83 | | **LIVINGAPARTMENTS\_MEDI** | | |  |
| 84 | | **LIVINGAREA\_MEDI** | | |  |
| 85 | | **NONLIVINGAPARTMENTS\_MEDI** | | |  |
| 86 | | **NONLIVINGAREA\_MEDI** | | |  |
| 87 | | **FONDKAPREMONT\_MODE.** | | |  |
| 88 | | **HOUSETYPE\_MODE.** | | |  |
| 89 | | **TOTALAREA\_MODE.** | | |  |
| 90 | | **WALLSMATERIAL\_MODE.** | | |  |
| 91 | | **EMERGENCYSTATE\_MODE** | | |  |
| 92 | | **OBS\_30\_CNT\_SOCIAL\_CIRCLE** | | | Cuántas veces ha registrado mora más de 30 días su entorno. |
| 93 | | **DEF\_30\_CNT\_SOCIAL\_CIRCLE** | | | Cuántas veces ha registrado mora más de 30 días su entorno. |
| 94 | | **OBS\_60\_CNT\_SOCIAL\_CIRCLE** | | | Cuántas veces ha registrado mora más de 60 días su entorno. |
| 95 | | **DEF\_60\_CNT\_SOCIAL\_CIRCLE** | | | Cuántas veces ha registrado mora más de 60 días su entorno. |
| 96 | | **DAYS\_LAST\_PHONE\_CHANGE** | | | Hace cuantos días antes de la postulación cambia número de teléfono. |
| 97 | | **FLAG\_DOCUMENT\_2** | | | Indicador documento 2. |
| 98 | | **FLAG\_DOCUMENT\_3** | | | Indicador documento 3. |
| 99 | | **FLAG\_DOCUMENT\_4** | | | Indicador documento 4. |
| 100 | | **FLAG\_DOCUMENT\_5** | | | Indicador documento 5. |
| 101 | | **FLAG\_DOCUMENT\_6** | | | Indicador documento 6. |
| 102 | | **FLAG\_DOCUMENT\_7** | | | Indicador documento 7. |
| 103 | | **FLAG\_DOCUMENT\_8** | | | Indicador documento 8. |
| 104 | | **FLAG\_DOCUMENT\_9** | | | Indicador documento 9. |
| 105 | | **FLAG\_DOCUMENT\_10** | | | Indicador documento 10. |
| 106 | | **FLAG\_DOCUMENT\_11** | | | Indicador documento 11. |
| 107 | | **FLAG\_DOCUMENT\_12** | | | Indicador documento 12. |
| 108 | | **FLAG\_DOCUMENT\_13** | | | Indicador documento 13. |
| 109 | | **FLAG\_DOCUMENT\_14** | | | Indicador documento 14. |
| 110 | | **FLAG\_DOCUMENT\_15** | | | Indicador documento 15. |
| 111 | | **FLAG\_DOCUMENT\_16** | | | Indicador documento 16. |
| 112 | | **FLAG\_DOCUMENT\_17** | | | Indicador documento 17. |
| 113 | | **FLAG\_DOCUMENT\_18** | | | Indicador documento 18. |
| 114 | | **FLAG\_DOCUMENT\_19** | | | Indicador documento 19. |
| 115 | | **FLAG\_DOCUMENT\_20** | | | Indicador documento 20. |
| 116 | | **FLAG\_DOCUMENT\_21** | | | Indicador documento 21. |
| 117 | | **AMT\_REQ\_CREDIT\_BUREAU\_HOUR** | | | Cantidad de consultas sobre el cliente al buró de crédito. Una hora antes de la postulación. |
| 118 | | **AMT\_REQ\_CREDIT\_BUREAU\_DAY** | | | Cantidad de consultas sobre el cliente al buró de crédito. Un día antes de la postulación. |
| 119 | | **AMT\_REQ\_CREDIT\_BUREAU\_WEEK** | | | Cantidad de consultas sobre el cliente al buró de crédito. Una semana antes de la postulación |
| 120 | | **AMT\_REQ\_CREDIT\_BUREAU\_MON** | | | Cantidad de consultas sobre el cliente al buró de crédito. Un mes antes de la postulación. |

## Detalle de nulos

|  |  |  |
| --- | --- | --- |
| Atributo | Total Nulos | Porcentaje |
| COMMONAREA\_AVG | 195236.000000 | 70.170218 |
| COMMONAREA\_MODE | 195236.000000 | 70.170218 |
| COMMONAREA\_MEDI | 195236.000000 | 70.170218 |
| NONLIVINGAPARTMENTS\_AVG | 194054.000000 | 69.745392 |
| NONLIVINGAPARTMENTS\_MODE | 194054.000000 | 69.745392 |
| NONLIVINGAPARTMENTS\_MEDI | 194054.000000 | 69.745392 |
| FONDKAPREMONT\_MODE | 191063.000000 | 68.670390 |
| LIVINGAPARTMENTS\_MEDI | 191063.000000 | 68.670390 |
| LIVINGAPARTMENTS\_AVG | 191063.000000 | 68.670390 |
| LIVINGAPARTMENTS\_MODE | 191063.000000 | 68.670390 |
| FLOORSMIN\_AVG | 189714.000000 | 68.185543 |
| FLOORSMIN\_MODE | 189714.000000 | 68.185543 |
| FLOORSMIN\_MEDI | 189714.000000 | 68.185543 |
| YEARS\_BUILD\_AVG | 185991.000000 | 66.847451 |
| YEARS\_BUILD\_MEDI | 185991.000000 | 66.847451 |
| YEARS\_BUILD\_MODE | 185991.000000 | 66.847451 |
| OWN\_CAR\_AGE | 183780.000000 | 66.052790 |
| LANDAREA\_MEDI | 166332.000000 | 59.781765 |
| LANDAREA\_AVG | 166332.000000 | 59.781765 |
| LANDAREA\_MODE | 166332.000000 | 59.781765 |
| BASEMENTAREA\_MEDI | 163954.000000 | 58.927082 |
| BASEMENTAREA\_AVG | 163954.000000 | 58.927082 |
| BASEMENTAREA\_MODE | 163954.000000 | 58.927082 |
| EXT\_SOURCE\_1 | 158015.000000 | 56.792533 |
| NONLIVINGAREA\_AVG | 154702.000000 | 55.601800 |
| NONLIVINGAREA\_MODE | 154702.000000 | 55.601800 |
| NONLIVINGAREA\_MEDI | 154702.000000 | 55.601800 |
| ELEVATORS\_MODE | 149491.000000 | 53.728902 |
| ELEVATORS\_AVG | 149491.000000 | 53.728902 |
| ELEVATORS\_MEDI | 149491.000000 | 53.728902 |
| WALLSMATERIAL\_MODE | 142642.000000 | 51.267288 |
| APARTMENTS\_AVG | 142408.000000 | 51.183185 |
| APARTMENTS\_MEDI | 142408.000000 | 51.183185 |
| APARTMENTS\_MODE | 142408.000000 | 51.183185 |
| ENTRANCES\_AVG | 141322.000000 | 50.792864 |
| ENTRANCES\_MEDI | 141322.000000 | 50.792864 |

|  |  |  |
| --- | --- | --- |
| Atributo | Total Nulos | Porcentaje |
| ENTRANCES\_MODE | 141322.000000 | 50.792864 |
| HOUSETYPE\_MODE | 140831.000000 | 50.616392 |
| LIVINGAREA\_MEDI | 140817.000000 | 50.611360 |
| LIVINGAREA\_MODE | 140817.000000 | 50.611360 |
| LIVINGAREA\_AVG | 140817.000000 | 50.611360 |
| FLOORSMAX\_MEDI | 139669.000000 | 50.198755 |
| FLOORSMAX\_MODE | 139669.000000 | 50.198755 |
| FLOORSMAX\_AVG | 139669.000000 | 50.198755 |
| YEARS\_BEGINEXPLUATATION\_MEDI | 136938.000000 | 49.217200 |
| YEARS\_BEGINEXPLUATATION\_MODE | 136938.000000 | 49.217200 |
| YEARS\_BEGINEXPLUATATION\_AVG | 136938.000000 | 49.217200 |
| TOTALAREA\_MODE | 135514.000000 | 48.705397 |
| EMERGENCYSTATE\_MODE | 133109.000000 | 47.841010 |
| OCCUPATION\_TYPE | 88800.000000 | 31.915811 |
| EXT\_SOURCE\_3 | 54658.000000 | 19.644757 |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK | 37239.000000 | 13.384154 |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR | 37239.000000 | 13.384154 |
| AMT\_REQ\_CREDIT\_BUREAU\_MON | 37239.000000 | 13.384154 |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT | 37239.000000 | 13.384154 |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY | 37239.000000 | 13.384154 |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | 37239.000000 | 13.384154 |
| NAME\_TYPE\_SUITE | 1007.000000 | 0.3619280 |
| EXT\_SOURCE\_2 | 599.000000 | 0.2152880 |
| AMT\_ANNUITY | 12.000000 | 0.0043130 |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | 1.000000 | 0.0003590 |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | 1.000000 | 0.0003590 |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | 1.000000 | 0.0003590 |
| DAYS\_LAST\_PHONE\_CHANGE | 1.000000 | 0.0003590 |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | 1.000000 | 0.0003590 |

## Gráficos de análisis univariado

|  |  |  |
| --- | --- | --- |
| **Gráficos de atributos numéricos** | | |
| Interfaz de usuario gráfica  Descripción generada automáticamente con confianza media | **AMT\_ICOME\_TOTAL: Ingreso total del cliente** | |
| Gráfico, Gráfico de dispersión  Descripción generada automáticamente | Gráfico, Histograma  Descripción generada automáticamente |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **AMT\_CREDIT: Cantidad total del préstamo realizado** | |
| Gráfico, Histograma  Descripción generada automáticamente | Gráfico, Gráfico de líneas, Histograma  Descripción generada automáticamente |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **AMT\_ANNUITY: Anualidad del préstamo** | |
| Gráfico, Histograma  Descripción generada automáticamente | Gráfico, Gráfico de líneas  Descripción generada automáticamente |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **AMT\_GOODS\_PRICE: Precio de los bienes que se comprará con el préstamo** | |
| Gráfico, Gráfico de líneas  Descripción generada automáticamente | Gráfico, Histograma  Descripción generada automáticamente |

|  |  |  |
| --- | --- | --- |
| **Gráficos de atributos numéricos** | | |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **REGION\_POPULATION\_RELATIVE: Población donde vive el cliente** | |
| Gráfico, Histograma  Descripción generada automáticamente | |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **AGE: Edad del cliente** | |
| Gráfico, Histograma  Descripción generada automáticamente | |
| Tabla  Descripción generada automáticamente con confianza baja | **DAYS\_EMPLOYED: Cantidad de días trabajados previos a la postulación** | |
| Tabla  Descripción generada automáticamente | Gráfico  Descripción generada automáticamente |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **DAYS\_REGISTRATION: Cantidad de días previos a la última modificación de los registros del cliente previos a la postulación.** | |
| Gráfico, Histograma  Descripción generada automáticamente | Gráfico  Descripción generada automáticamente |

|  |  |  |
| --- | --- | --- |
| **Gráficos de atributos numéricos** | | |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **DAYS\_ID\_PUBLISH: Cantidad de días previos a la modificación de su documento de identificación con el cual realizo la postulación de préstamo** | |
| Gráfico, Histograma  Descripción generada automáticamente | Gráfico  Descripción generada automáticamente |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **DAYS\_LAST\_PRHONE\_CHANGE: Hace cuántos días antes de la postulación cambia número de teléfono** | |
| Gráfico, Histograma  Descripción generada automáticamente | Gráfico  Descripción generada automáticamente |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **EXT\_SOURCE\_2: Puntaje normalizado de fuente externa 2** | |
| Gráfico, Histograma  Descripción generada automáticamente | |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente | **EXT\_SOURCE\_3: Puntaje normalizado de fuente externa 3** | |
| Gráfico, Histograma  Descripción generada automáticamente | |

|  |  |  |
| --- | --- | --- |
| **Gráficos de atributos numéricos** | | |
| **FLAG\_PHONE: Da un teléfono de contacto** | **FLAG\_EMP\_PHONE: Da un teléfono de trabajo de contacto** | **FLAG\_WORK\_PHONE: Da un teléfono de hogar de contacto** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **LIVE\_CITY\_NOT\_WORK\_CITY: Dirección de contacto del cliente no concuerda con la dirección de trabajó** | **REG\_CITY\_NOT\_LIVE\_CITY: Dirección permanente no concuerda con la dirección de contacto** | **REG\_CITY\_NOT\_WORK\_CITY:**  **Dirección permanente no concuerda con la dirección del trabajo** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente |  |  |
| **REGION\_RATING\_CLIENT: Evaluación inerna sobre la region donde vive el cliente** | **DEF\_30\_CNT\_SOCIAL\_CIRCLE: Cuántas veces ha registrado mora más de 30 días su entorno** | **DEF\_60\_CNT\_SOCIAL\_CIRCLE: Cuántas veces ha registrado mora más de 60 días su entorno** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |

|  |  |
| --- | --- |
| **Gráficos de atributos numéricos** | |
| **OBS\_30\_CNT\_SOCIAL\_CIRCLE: Cuántas veces ha registrado mora más de 30 días su entorno** | **OBS\_60\_CNT\_SOCIAL\_CIRCLE: Cuántas veces ha registrado mora más de 60 días su entorno** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **AMT\_REQ\_CREDIT\_BUREAU\_HOUR: Cantidad de consultas al buró de crédito. Una hora antes** | **AMT\_REQ\_CREDIT\_BUREAU\_DAY: Consulta al buró un día antes de la postulación** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **AMT\_REQ\_CREDIT\_BUREAU\_WEEK: Una semana antes** | **AMT\_REQ\_CREDIT\_BUREAU\_MON: Un mes antes** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **AMT\_REQ\_CREDIT\_BUREAU\_QRT: Tres meses antes** | **AMT\_REQ\_CREDIT\_BUREAU\_YEAR: Un año antes** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |

|  |  |  |
| --- | --- | --- |
| **Gráficos de atributos categóricos** | | |
| **CODE\_GENDER: Sexo del cliente** | **FLAG\_OWN\_CAR: Tiene auto** | **FLAG\_OWN\_REALTY: Tiene casa** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **FLAG\_MOBIL: Da un teléfono celular de contacto** | **FLAG\_CONT\_MOBILE: Se pudo contactar al cliente mediante celular** | **FLAG\_EMAIL: Da un email de contacto el cliente** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **REGION\_RATING\_CLIENT\_W\_CITY: Evaluación interna sobre la ciudad donde vive el cliente** | **REG\_REGION\_NOT\_LIVE\_REGION: Dirección permanente del cliente no concuerda con la dirección de contacto** | **REG\_REGION\_NOT\_WORK\_REGION: Dirección permanente del cliente no concuerda con la dirección de trabajo** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **LIVE\_REGION\_NOT\_WORK\_REGION: Dirección de contacto del cliente no concuerda con la dirección de trabajo** | **WEEKDAY\_APPR\_PROCESS\_START: Día habil que pide el préstamo** | **HOUR\_APPR\_PROCESS\_START: Hora aproximada solicitud del préstamo** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |

|  |  |
| --- | --- |
| **Gráficos de atributos categóricas** | |
| **NAME\_INCOME\_TYPE: Tipo de ingreso por parte del cliente** | **Gráfico, Gráfico de barras, Gráfico de embudo  Descripción generada automáticamente** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **NAME\_EDUCATION\_TYPE: Máximo nivel educacional por parte del cliente** | **Gráfico, Gráfico de barras  Descripción generada automáticamente** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **NAME\_FAMILY\_STATUS: Situación familiar del cliente** | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **NAME\_HOUSING\_TYPE: Cuál es la situación habitacional del cliente** | **Gráfico, Gráfico de barras  Descripción generada automáticamente** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente |

|  |  |
| --- | --- |
| **Gráficos de atributos categóricas** | |
| **OOCUPATION\_TYPE: Distribución de la profesión del cliente** | **Gráfico, Gráfico de barras  Descripción generada automáticamente** |
| Gráfico, Gráfico de barras, Gráfico de embudo  Descripción generada automáticamente |
| **ORGANIZATION\_TYPE: Tipo de organización donde trabaja el cliente** | |
| Interfaz de usuario gráfica, Aplicación  Descripción generada automáticamente | |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | |

|  |  |  |
| --- | --- | --- |
| **Gráficos de atributos categóricos** | | |
| **FLAG\_DOCUMENT\_2: Indicador documento 2** | **FLAG\_DOCUMENT\_3: Indicador documento 3** | **FLAG\_DOCUMENT\_4: Indicador documento 4** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **FLAG\_DOCUMENT\_5: Indicador documento 5** | **FLAG\_DOCUMENT\_6: Indicador documento 6** | **FLAG\_DOCUMENT\_7: Indicador documento 7** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **FLAG\_DOCUMENT\_8: Indicador documento 8** | **FLAG\_DOCUMENT\_9: Indicador documento 9** | **FLAG\_DOCUMENT\_10: Indicador documento 10** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **FLAG\_DOCUMENT\_11 Indicador documento 11** | **FLAG\_DOCUMENT\_12: Indicador documento 12** | **FLAG\_DOCUMENT\_13: Indicador documento 13** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |

|  |  |  |
| --- | --- | --- |
| **Gráficos de atributos categóricos** | | |
| **FLAG\_DOCUMENT\_14: Indicador documento 14** | **FLAG\_DOCUMENT\_15: Indicador documento 15** | **FLAG\_DOCUMENT\_16: Indicador documento 16** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **FLAG\_DOCUMENT\_17: Indicador documento 17** | **FLAG\_DOCUMENT\_18: Indicador documento 18** | **FLAG\_DOCUMENT\_19: Indicador documento 19** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |
| **FLAG\_DOCUMENT\_20: Indicador documento 20** | **FLAG\_DOCUMENT\_21: Indicador documento 21** |  |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |  |

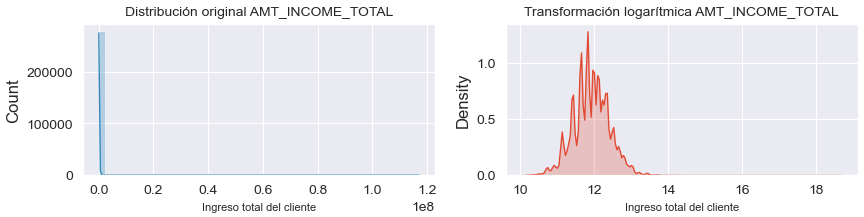
|  |  |
| --- | --- |
| **Gráficos de atributos numéricos** | |
| **CNT\_CHILDREN: Cantidad de hijos por parte del cliente** | **CNT\_FAM\_MEMBERS: Cuantos miembros familiares tiene el cliente** |
| Gráfico, Gráfico de barras  Descripción generada automáticamente | Gráfico, Gráfico de barras  Descripción generada automáticamente |

## Categorizaciones

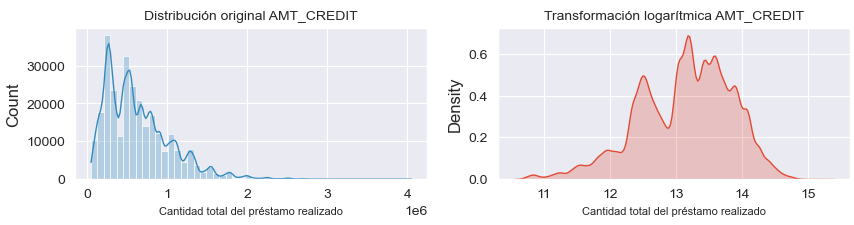
**Transformación de Variables:**

1. Se transformó la edad en años de la variable DAYS\_BIRTH a valores positivos, según lo siguiente: *AGE = DAYS\_BIRTH \*-1 / 365*

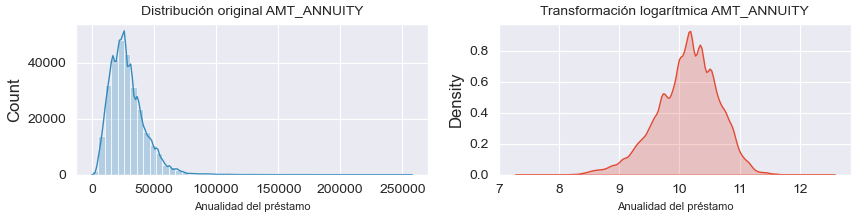
2. Se transformó la variable AMT\_INCOME\_TOTAL a logaritmo que indica el ingreso total del cliente, presentando una mejora más cercana a la distribución normal.



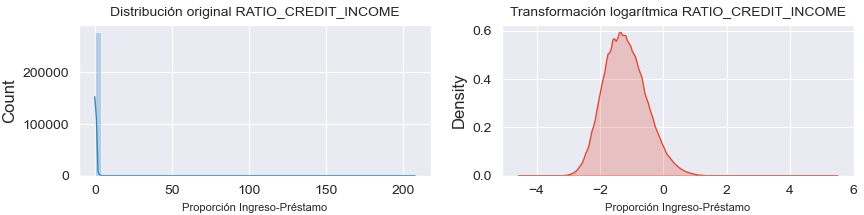
3. Se transformó la variable AMT\_CREDIT a logaritmo que indica la cantidad total del préstamo realizado, presentando una mejora más cercana a la distribución normal.



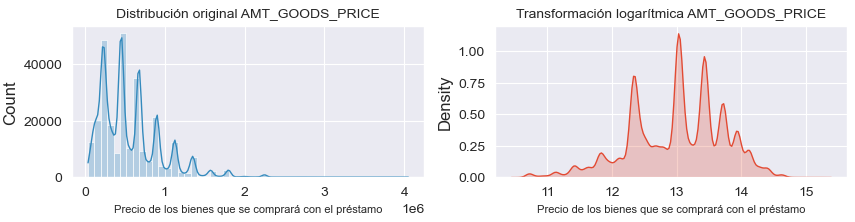
4. Se transformó la variable AMT\_ANNUITY a logaritmo que indica la anualidad del préstamo, presentando una mejora más cercana a la distribución normal.



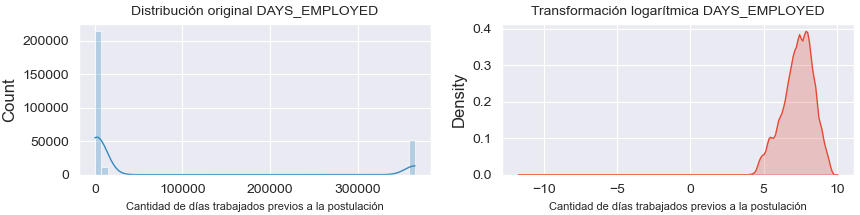
5. Se transformó la variable RATIO\_CREDIT\_INCOME a logaritmo que indica proporción ingreso-préstamo, presentando una mejora más cercana a la distribución normal.



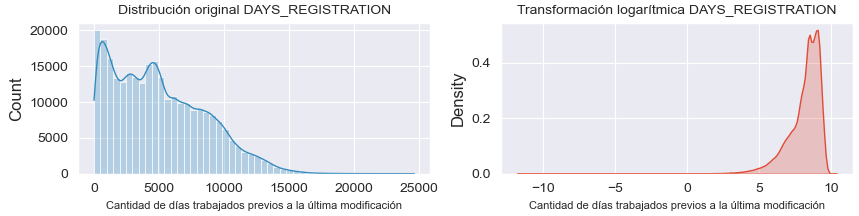
6. Se transformó la variable AMT\_GOODS\_PRICE a logaritmo que indica precio de los bienes que se comprará con el préstamo, presentando una mejora más cercana a la distribución normal.



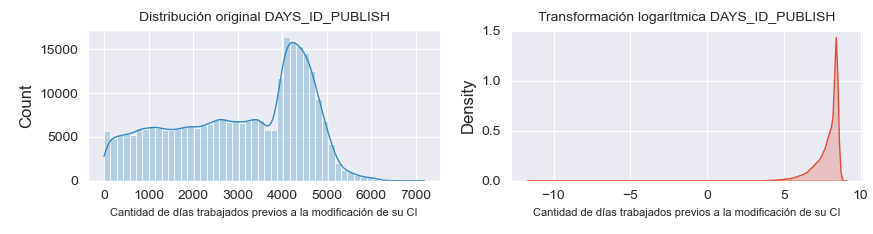
7. Se transformó la variable DAYS\_EMPLOYED a valores positivos y luego a logaritmo que indica la cantidad de días trabajados previos a la postulación, presentando una mejora más cercana a la distribución normal.



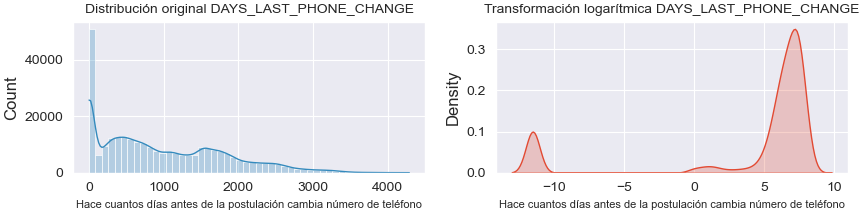
8. Se transformó la variable DAYS\_REGISTRATION a valores positivos y luego a logaritmo que indica la cantidad de días previos a la última modificación de los registros del cliente previos a la postulación, presentando una mejora más cercana a la distribución normal.



9. Se transformó la variable DAYS\_ID\_PUBLISH a valores positivos y luego a logaritmo que indica la cantidad de días previos a la modificación de su Cedula de Identidad con el cual realizó la postulación, presentando una mejora más cercana a la distribución normal.



10. Se transformó la variable DAYS\_LAST\_PHONE\_CHANGE a logaritmo que indica a hace cuantos días antes de la postulación cambia número de teléfono, presentando una mejora más cercana a la distribución normal.



**Creación de Variables:**

1. Se creó variable RATIO\_CREDIT\_INCOME que indica la proporción ingreso-préstamo en relación a las variables AMT\_INCOME\_TOTAL y AMT\_CREDIT, según lo siguiente: *RATIO\_CREDIT\_INCOME = AMT\_INCOME\_TOTAL / AMT\_CREDIT*

**Recodificación de Variables:**

1. Se recodificó la variable NAME\_INCOME\_TYPE correspondiente al ingreso por parte del cliente. Se disminuyó de 7 a 4 categorías según análisis de la siguiente tabla:

|  |  |
| --- | --- |
| NAME\_INCOME\_TYPE | |
| Variable inicial | **Recodificación** |
| Commercial associate | Commercial |
| State servant | Government |
| Maternity leave | Others |
| Pensioner |
| Student |
| Unemployed |
| Working | Working |

1. Se recodificó la variable NAME\_EDUCATION\_TYPE correspondiente al nivel de estudio del cliente. Se disminuyó de 5 a 3 categorías según análisis de la siguiente tabla:

|  |  |
| --- | --- |
| NAME\_INCOME\_TYPE | |
| Variable inicial | **Recodificación** |
| Lower secondary | School |
| Incomplete higher | Secondary |
| Secondary/secondary special |
| Academic degree | University |
| Higher education |

3.-Se recodificó la variable NAME\_FAMILY\_STATUS correspondiente a situación familiar del cliente. Se disminuyó de 5 a 2 categorías según análisis de la siguiente tabla:

|  |  |
| --- | --- |
| NAME\_FAMILY\_STATUS | |
| Variable inicial | **Recodificación** |
| Civil marriage | Married |
| Married |
| Separated | Not\_married |
| Single/not married |
| Widow |

4.- Se recodificó la variable NAME\_HOUSING\_TYPE correspondiente a si el cliente tiene casa o departamento. Se disminuyó de 6 a 3 categorías según análisis de la siguiente tabla:

|  |  |
| --- | --- |
| NAME\_HOUSING\_TYPE | |
| Variable inicial | **Recodificación** |
| Co-op apartment | Others |
| House/apartment | House\_aparment |
| Municipal apartment | Others |
| Office apartment | Others |
| Rented apartment | With\_parents\_or\_rented |
| With parents | With\_parents\_or\_rented |

5.- Se recodificó la variable OCCUPATION\_TYPE correspondiente a la profesión del cliente. Se disminuyó de 18 a 5 categorías según análisis de la siguiente tabla:

|  |  |
| --- | --- |
| OCCUPATION\_TYPE | |
| Variable inicial | **Recodificación** |
| Core staff | Occupation\_type\_G1 |
| High skill tech staff |
| HR staff |
| IT staff |
| Managers |
| Medicine staff |
| Private service staff |
| Realty agents |
| Secretaries |
| Cleaning staff | Occupation\_type\_G2 |
| Sales staff |
| Cooking staff | Occupation\_type\_G3 |
| Laborers |
| Security staff |
| Drivers | Occupation\_type\_G4 |
| Low-skill Laborers |
| Waiters/barmen staff |
| Accountants | Occupation\_type\_G5 |

6.- Se recodificó la variable ORGANIZATION\_TYPE correspondiente al tipo de organización donde trabaja el cliente. Se disminuyó de 58 a 8 categorías según análisis de la siguiente tabla:

|  |  |
| --- | --- |
| ORGANIZATION\_TYPE | |
| Variable inicial | **Recodificación** |
| Construction | Organization\_type\_G1 |
| Realtor |
| Restaurant |
| Agriculture | Organization\_type\_G2 |
| Cleaning |
| Security |
| Self-employed |
| Transport |
| Business | Organization\_type\_G3 |
| Entity |
| Trade |
| Advertising | Organization\_type\_G4 |
| Industry |
| Mobile |
| Housing | Organization\_type\_G5 |
| Legal Services |
| Other |
| Electricity | Organization\_type\_G6 |
| Emergency |
| Government |
| Kindergarten |
| Medicine |
| Services |
| Telecom |
| Hotel | Organization\_type\_G7 |
| Insurance |
| Religion |
| School |
| Bank | Organization\_type\_G8 |
| Culture |
| Military |
| Police |
| Security |
| Ministries |
| University y XNA |

## Gráficos de análisis bivariado

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente

Interfaz de usuario gráfica, Gráfico, Aplicación

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Histograma

Descripción generada automáticamente

Gráfico, Gráfico radial

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente con confianza media

## Marco teórico de algoritmos utilizados

**Gradient Boosted regression trees**

Este método combina múltiples arboles de decisión para crear un mejor modelo, este tipo de algoritmo puede ser utilizado para regresión y clasificación. En contraste con el enfoque de Random Forest, trabaja mediante la construcción de árboles en serie, donde cada árbol trata de corregir los errores del anterior. Por defecto, hay no hay aleatorización en árboles de regresión potenciados por gradiente; en cambio, fuerte pre-poda se usa. Los árboles potenciados por gradiente a menudo usan árboles muy poco profundos, de una profundidad de uno a cinco, lo que hace que el modelo sea más pequeño en términos de memoria y hace que las predicciones sean más rápidas.

**GAM**

Es un modelo lineal generalizado en el que la variable de respuesta lineal depende linealmente de funciones suaves desconocidas de algunas variables predictoras, y el interés se centra en la inferencia sobre estas funciones suaves.

Los GAM fueron desarrollados originalmente por Trevor Hastie y Robert Tibshirani para combinar propiedades de modelos lineales generalizados con modelos aditivos.

El modelo relaciona una variable de respuesta univariada, Y, con algunas variables predictoras, xi. Se especifica una distribución de familia exponencial para Y (por ejemplo, distribuciones normales, binomiales o de Poisson) junto con una función de enlace g (por ejemplo, las funciones de identidad o logarítmica) que relacionan el valor esperado de Y con las variables predictoras a través de una estructura como:



**Regresión Logística**

Este método es habitualmente usado para abordar problemas de clasificación, en donde opera como un clasificador binario, para lo cual crea pesos para cada uno de los atributos que participan del cálculo, la ecuación se representa de la siguiente forma:

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

**Análisis discriminante**

Este método de clasificación aprende a discriminar entre las diferentes clases definiendo hiperplanos en donde proyecta los datos, así también es una buena técnica para reducir dimensionalidad.

**Adaboost**

Este método presenta una metodología iterativa en donde calculo pesos y luego realiza predicciones sobre los datos de entrenamiento y luego se vuelve a repetir el procedimiento iterativamente minimizando el error.

Función de tasa de error ponderada:

Texto

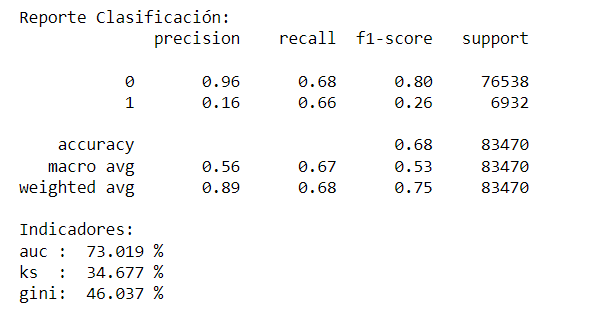
Descripción generada automáticamente

**Random Forest**

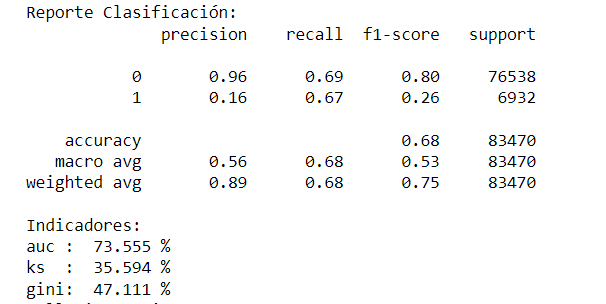
Es un ensamblaje de arboles de decisiones, generalmente entrenados vía un método de bagging, para lo cual busca los mejores atributos que permiten separar la clase objetivo.

## Reportes de resultados de los modelos

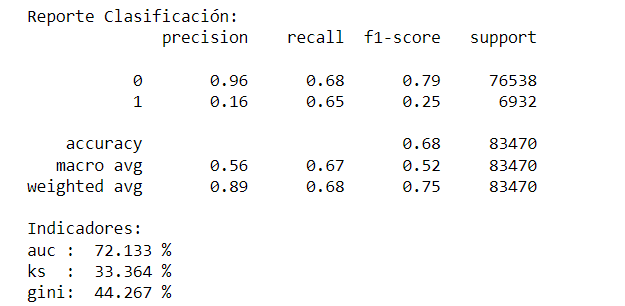
**Regresión Logística:**

****

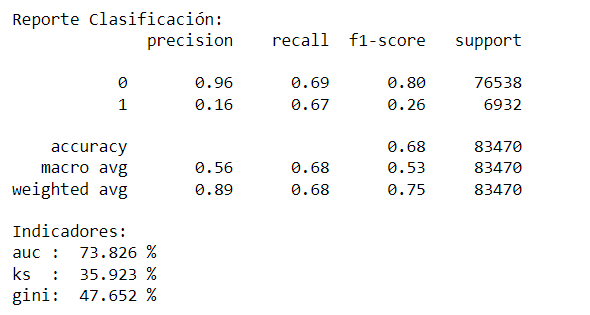
**LogisticGAM**

****

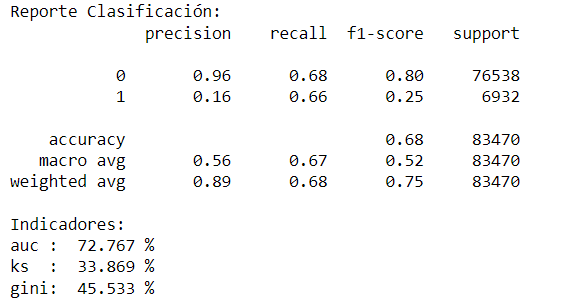
**RandomForest**

****

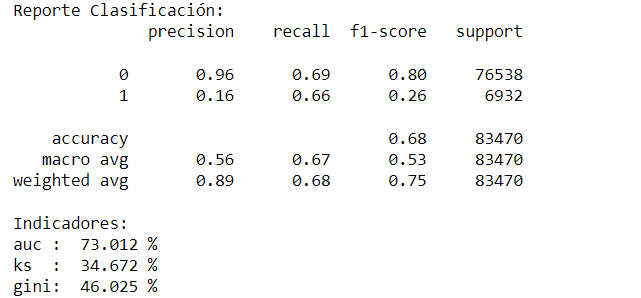
**GradientBoosting**



**AdaBoost:**



**LinearDiscriminantAnalysis:**



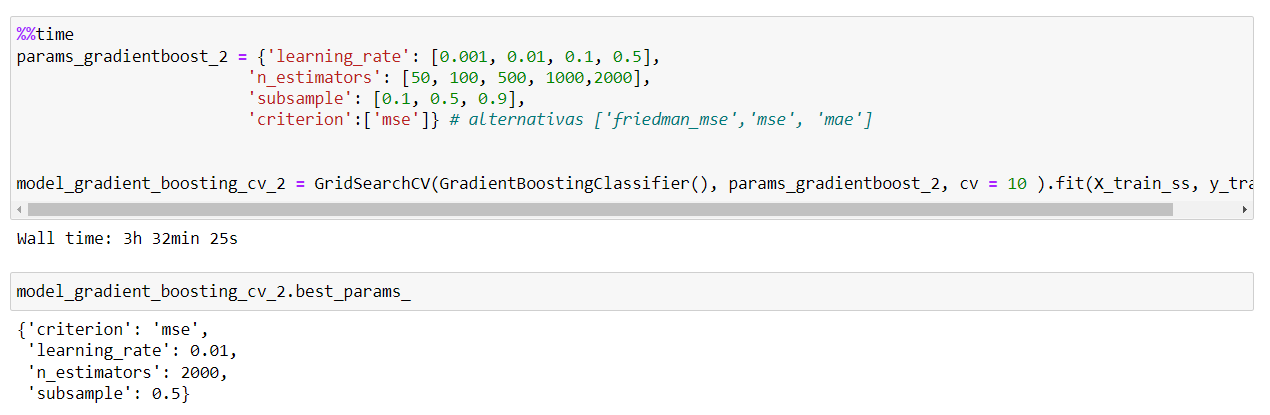
## Búsqueda hiperparámetros modelos seleccionados.

## Regresión Logística.



## Grandient boosting.

Se realiza búsqueda a continuación:



## Bibliografía

[1] Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow (Aurélien Géron).

[2] Introduction to Machine Learning with Python. (Andreas C.Müller & Sarah Guido).