**MODELO ORIGEN NO BANK BPE**

Área Modelos y Metodologías

SQUAD BPExito

**INTERBANK**

**Control de Versiones**

|  |  |  |  |
| --- | --- | --- | --- |
| **Versión** | **Fecha** | **Realizado por** | **Actualización realizada** |
| 1 | 02/10/2023 | Mariluna Olarte | Diseño del modelo |
| 2 | 04/10/2023 | Mariluna Olarte | Actualización de gráficos |
| 3 | 06/10/2023 | Mariluna Olarte | Actualización de descripción de default |
| 4 | 13/10/2023 | Mariluna Olarte | Actualización del archivo, creación de variables, actualización de descripción de filtros. |
| 5 | 20/10/23 | Mariluna Olarte | Se modificó el nombre de la variable n\_meses\_anofabricacion\_antiguov2  por n\_meses\_anofabricacion\_nuevov2  Actualización del archivo creación de variables, descriptivo de variables(missing),actualización del cuadro de filtros |

Contenido

[1. INTRODUCCIÓN 4](#_Toc148714067)

[2. POBLACIÓN OBJETIVO 4](#_Toc148714068)

[3. RECOPILACIÓN Y GENERACIÓN DE INFORMACIÓN 6](#_Toc148714069)

[3.1 Fuentes de información requeridas 6](#_Toc148714070)

[3.2 Softwares utilizados 7](#_Toc148714071)

[4. DESCRIPCION DE PROCESOS 7](#_Toc148714072)

[4.1 OBTENCION DE LA MUESTRA 7](#_Toc148714073)

[4.2 FITLTROS APLICADOS SOBRE LA BASES DE DATOS INICIALES 8](#_Toc148714074)

[4.3 PERIODOS DE OBSERVACIÓN 9](#_Toc148714075)

[4.2 PERIODO DE COMPORTAMIENTO 10](#_Toc148714076)

[5. MARCA DE DEFAULT 11](#_Toc148714077)

[5.1 Análisis descriptivo 11](#_Toc148714078)

[5.2 Análisis temporal a nivel variable 12](#_Toc148714079)

[5.3 Creación de Variables 13](#_Toc148714080)

[6. ETAPA DE MODELAMIENTO 13](#_Toc148714081)

[6.1 Metodología de Modelamiento 13](#_Toc148714082)

[Gradient Boosting 13](#_Toc148714083)

[Extreme Gradient Boosting (XGBoost) 14](#_Toc148714084)

[6.2 Partición de la muestra 15](#_Toc148714085)

[6.3 Modelo Hiperpámetros del Modelo Extreme Gradient Boosting 16](#_Toc148714086)

[6.4 Selección de variables 17](#_Toc148714087)

[6.5 Análisis Bivariante/ Dependencia parcial con valores shap 18](#_Toc148714088)

[6.6 Resultados de Perfomance del modelo 19](#_Toc148714089)

[6.8 Sofware utilizado: 23](#_Toc148714090)

[7. ANEXOS 24](#_Toc148714091)

## INTRODUCCIÓN

Las empresas BPE (BPE definimos como aquellas empresas con niveles de ventas entre 180 000 y 3 millones soles anuales, con posibles excepciones) del rubro financiero necesitan poner foco a clientes potenciales que puedan adquirir un producto en el banco, por tanto, antes de efectuarse dicho proceso se necesita optar por un estudio sobre los clientes con el objetivo analizar la confiabilidad de estos en ser un cliente adecuado a nivel de riesgo para la entidad.

El siguiente documento contiene la descripción del modelo score de admisión BPE para no bancarizados. El cual corresponde a un modelo matemático que clasifica a un cliente según su perfil de riesgo, en alguna de las siguientes categorías: bueno o malo.

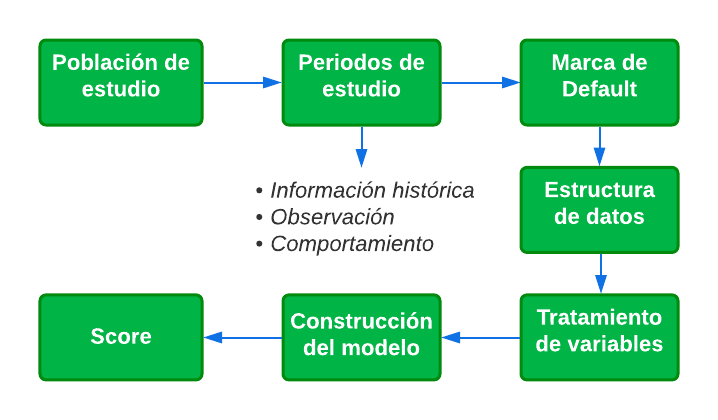
## POBLACIÓN OBJETIVO

La población objetivo son todos aquellos clientes que no están bancarizados, es decir, aquellos clientes que no presentan saldo negocio en los 12 últimos meses y en el mes de observación hace un desembolso mayor a 10mil.

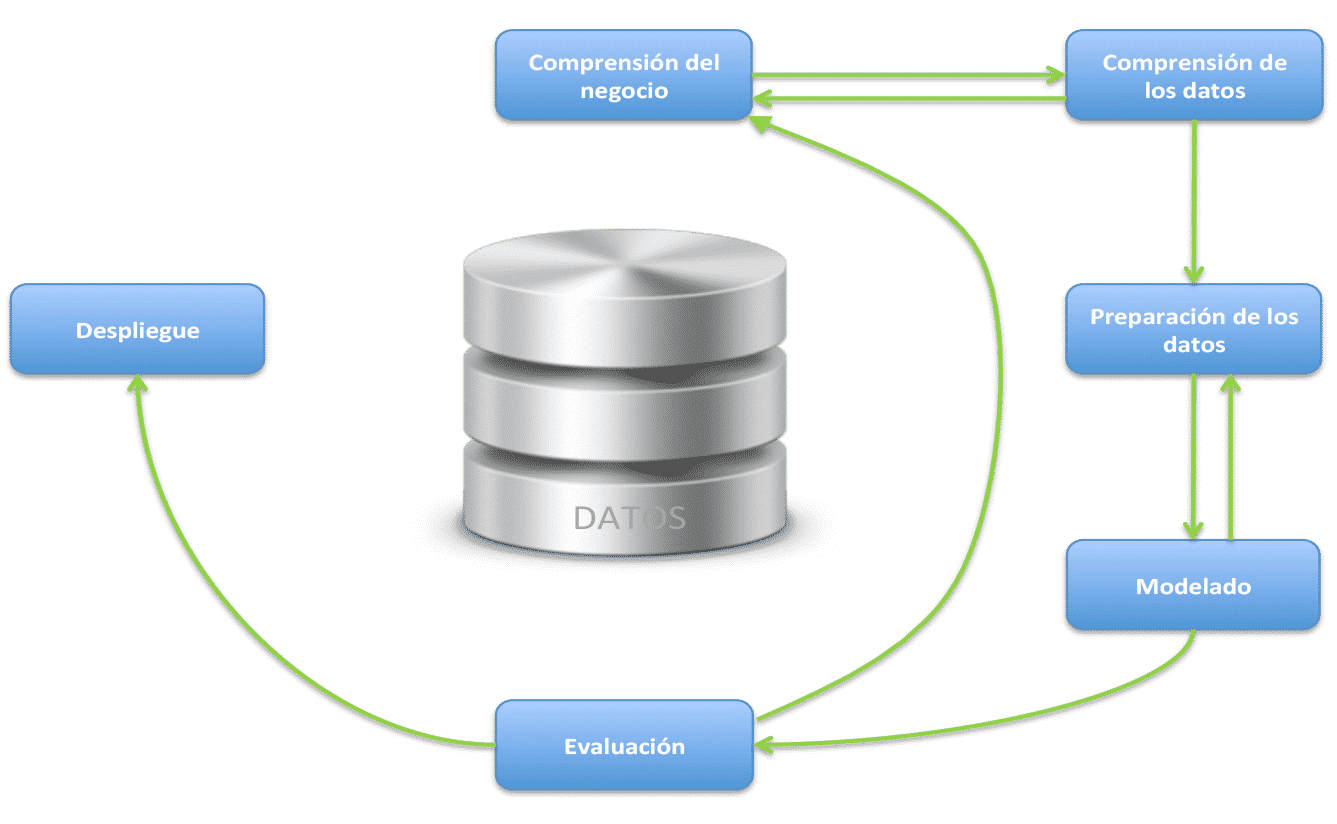
Adicional a esto se agregan a clientes 100% reactiva (*El programa Reactiva Perú, creado por el Estado mediante DL N°1455, está dirigido a las empresas afectadas por la emergencia sanitaria del COVID-19, y tiene como objetivo promover financiamiento de reposición de capital de trabajo de empresas que enfrentan pagos y obligaciones de corto plazo y así asegurar continuidad en la cadena de pagos*), es decir, clientes que en los 12 últimos meses presente solo saldo reactiva y que en el mes de observación hace un desembolso mayor al saldo reactiva.

Con el objetivo de evaluar el comportamiento de los clientes a futuro, se necesitará que los clientes en el periodo de observación no tengan un comportamiento negativo ni inconsistencias por lo que se optan por las siguientes exclusiones:

* Clientes refinanciados, reestructurados, judiciales y castigados en créditos negocios.
* Clientes con días mora en el periodo de observación mayor a 30 días.
* Cliente con Estado en SUNAT (Activo y Suspendido)



Metodología CRIPS-DM



## RECOPILACIÓN Y GENERACIÓN DE INFORMACIÓN

### Fuentes de información requeridas

La información para el desarrollo del modelo se obtiene de las siguientes fuentes mencionadas en el siguiente gráfico.

SUNAT

ADEX

RCC

SUNARP

DEMOGRAFICAS

RCC

DEMOGRAFICAS

IZIPAY

ENRICH

SENTINEL

* **Base RCC (Base Sistema Financiera):** Información de clientes en el sistema financiero.
* **Base SUNAT (Base Proveedor):** Base de contribuyentes para evaluar sus características. También contiene la información de representantes legales de los clientes, en su mayoría jurídicos.
* **Base ADEX (Base Proveedor):** Información sobre los clientes que realizan exportación o importación. Entre sus variables más resaltantes están montos y continentes que realizan la exportación o importación.
* **Base DEMOGRAFICAS (Base interna):** Contiene información general de los clientes, tomadas y consolidadas de diferentes fuentes.
* **Base SENTINEL (Base Proveedor):** Información del mercado, pueden ser los buros, consultoras, etc. Esta Data incluye geolocalización, profesiones, empresas, telecomunicaciones y empresas no reguladas.
* **Base ENRICH (Base Proveedor):** Información censal ENRICH sobre los clientes.
* **Base SUNARP (Base Proveedor):** Información de SUNARP de los clientes.
* **Base IZIPAY (Base interna):** Información de transacciones izipay.

**Rutas de las fuentes usadas:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | **ESQUEMA** | **TABLA** | **ACTUALIZACIÓN** |  |
|  | **e\_perm\_aws** | t\_fact\_report\_rcc\_rsk | MENSUAL |  |
|  | **e\_perm\_aws** | t\_mst\_princ\_sunat\_rsk t\_mst\_sec\_sunat\_rsk t\_mst\_rep\_legal\_sunat\_rsk | MENSUAL |  |
|  | **e\_perm\_aws** | t\_mst\_sunarp\_rsk | MENSUAL |  |
|  | **e\_perm\_aws** | mst\_vpc\_adex | MENSUAL |  |
|  | **e\_perm\_aws** | t\_rsk\_vars\_constr\_enrich | - |  |
|  | **e\_perm\_aws** | tbl\_per\_inf\_mdl | MENSUAL |  |
|  | **e\_perm\_aws** | t\_sentinel\_rsk | MENSUAL |  |
|  |  |  |  |  |

### Softwares utilizados

* Microsoft Excel, Word y Power Point usado para la construcción de la documentación
* Teradata y AWS (Athena) para la construcción de la base de desarrollo y validación
* AWS (SageMaker) para la construcción del modelo de desarrollo
* Python, como lenguaje para la construcción del modelo donde se usaron paquetes:
* Pandas= '1.3.4'
* Numpy= '1.20.3'
* Xgboost= '1.4.2'
* Shap= '0.40.0'
* Scorecardpy= '0.1.9.2'

## DESCRIPCION DE PROCESOS

### OBTENCION DE LA MUESTRA

Se ha considerado periodos actuales y donde la tasa de mora sea lo más similar entre meses para evitar el sobreajuste, se ha incluido a clientes reactiva, en donde el saldo desembolsado sea mayor al saldo reactiva en el mes de alta.

Los períodos mostrados en los siguientes gráficos hacen referencia a la fecha de alta (desembolso).

### FITLTROS APLICADOS SOBRE LA BASES DE DATOS INICIALES

* *Solo presentan desembolsos >10,000 (flag\_flag\_desembolso=1):*

**Motivo**: Son clientes que no presentan ningún tipo de deuda en créditos negocios en el periodo de observación.

* *Malos en el origen (flag\_malo\_orig=1):*

**Motivo**: Son clientes que son malos al momento de la observación por lo que no hay nada que aprender de ellos, estos clientes presentan un perfil riesgoso, y el modelo no bancarizado necesita aprender de clientes que no presentan deterioro en el sistema financiero. Por ello son clientes que tienen días de atraso mayor a cero, son castigados, refinanciados, reestructurados y judiciales en el mes de observación.

* *Duplicados (flag\_duplicados=1):*

**Motivo**: Son clientes que se duplican dado que cumplen con ambas características, es decir es un no bancarizado y a las vez es un reactiva, por ello solo nos quedamos con los clientes reactivas.

* *Insuficientes (flag\_insuf=1):*

**Motivo**: Se define como insuficiente a aquellos clientes que tienen menos de 6 meses de performance (saldo).

* *Indeterminados (flag\_ind=1):*

**Motivo**: Son clientes que no podemos estar seguros de que sean malos o buenos, debido a la alta tasa de recuperación de impago (de 31 a 90 días) que existe, se define como indeterminados a los clientes que alcanzan más de 30 días, pero no los 91 días y a la vez no sean clientes castigados, refinanciados, refinanciados reestructurados ni judiciales en los próximos 12 meses.

* *Estado Activo y Suspendido (flag\_activ\_susp=1):*

**Motivo**: Son clientes diferente al estado de SUNAT activo y suspendido no son evaluados.

* *Perfil BPE (flg\_perfil\_bpe=1):*

**Motivo**: Los clientes que no cumplan con las siguientes características no son evaluados:

Persona jurídica o persona natural con negocio, mayor de 25 años y menor de 75 años con clasificación normal en el últimos mes y hasta CPP en los últimos 12 meses.

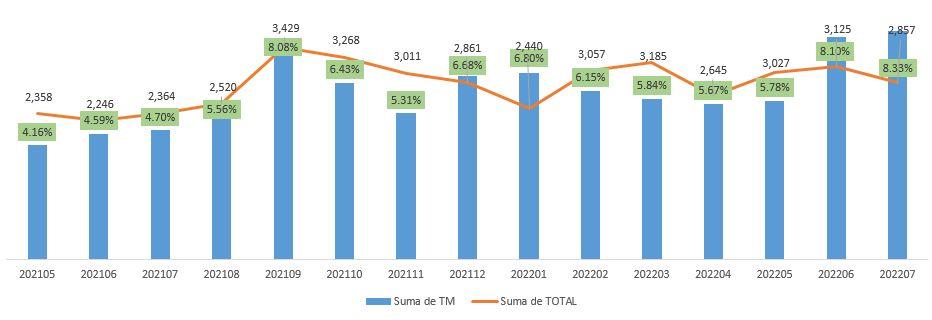
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Exclusiones y filtros Target** | **Periodo** | | | |
| **202105-202207** | | | |
| **No Bank** | **Filtrados** | **Reactiva** | **Filtrados** |
| Base Inicial | 2,662,082 |  | 33,220 |  |
| Desembolso>10,000 | 175,717 | 2,486,365 | 15,060 | 18,160 |
| malo en el origen | 171,821 | 3,896 | 12,623 | 2,437 |
| Base consolidada (no Bank + reactiva)(sin duplicados) | 184,444 |  | - | - |
| Duplicados | 184,376 | 68 |  |  |
| Insuficientes | 172,806 | 11,570 | - | - |
| indeterminados | 168,359 | 4,447 | - | - |
| Activo y Suspendido | 110,176 | 58,183 | - | - |
| Perfil BPE | 42,393 |  |  |  |
| Base final | 42,393 |  | - | - |

### PERIODOS DE OBSERVACIÓN

La muestra se desarrolló con una ventana de observación conformada por 15 cosechas tomando como periodos desde May21 a Jul22.

A continuación una vista global de la muestra de desarrollo:

|  |  |  |
| --- | --- | --- |
| **codmes** | **TOTAL** | **TM** |
| **202105** | **2,358** | **4.16%** |
| **202106** | **2,246** | **4.59%** |
| **202107** | **2,364** | **4.70%** |
| **202108** | **2,520** | **5.56%** |
| **202109** | **3,429** | **8.08%** |
| **202110** | **3,268** | **6.43%** |
| **202111** | **3,011** | **5.31%** |
| **202112** | **2,861** | **6.68%** |
| **202201** | **2,440** | **6.80%** |
| **202202** | **3,057** | **6.15%** |
| **202203** | **3,185** | **5.84%** |
| **202204** | **2,645** | **5.67%** |
| **202205** | **3,027** | **5.78%** |
| **202206** | **3,125** | **8.10%** |
| **202207** | **2,857** | **8.33%** |



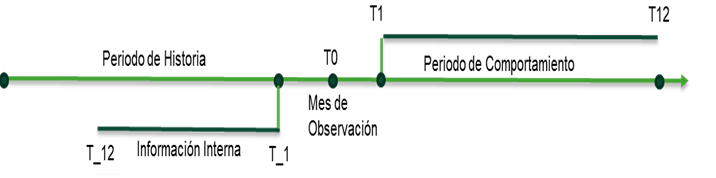
### PERIODO DE COMPORTAMIENTO

Los periodos de comportamiento hacen referencia a los meses en que se observa el desempeño del cliente, posteriores a un mes inicial de observación. Esto corresponde a los de 15 meses siguientes respecto a cada mes de observación, que comprenden los periodos de Mayo-2021 hasta Julio-2022.

Son en estos periodos en donde se evaluará el comportamiento de los clientes, determinándose una marca de comportamiento negativo, denominado Default, el cual será descrito posteriormente.

* 1. **PERIODO DE INFORMACIÓN HISTÓRICA**

Los periodos de información histórica para la construcción del modelo, hacen referencia a los periodos en donde se recogerán información de los clientes. Particularmente en este modelo, estos periodos comprenden desde un periodo precedente al mes de observación hasta 12 periodos como máximo.



## MARCA DE DEFAULT

La marca de incumplimiento, denominado Default, hace referencia al comportamiento negativo que un cliente presenta en su periodo de comportamiento. Esta marca representa como malo, en un horizonte de 12 meses, aquel cliente (natural o jurídica) que presenta alguna de las siguientes condiciones:

* Si posee uno o más créditos refinanciados
* Si posee uno o más créditos reestructurados
* Si posee uno o más créditos castigados
* Si posee uno o más créditos judiciales
* Si su deuda con atraso superior a 90 días

Estos días de mora, refinanciando y castigados considera todos los productos de saldo negocio empresarial a nivel sistema financiero.

El Default es denotado por el valor de 1 (clientes malos), y los clientes que no presentan comportamiento negativo son denotados con 0 (clientes buenos)

También se define la tasa de Default, la cual mide la proporción de clientes malos sobre el total de clientes

### Análisis descriptivo

Para tener un conocimiento a rasgos generales de la información que se dispone y, al mismo tiempo, detectar posibles errores y falta de información, se realiza un “análisis descriptivo”.

Se diferencian dos tipos de análisis en función de si la variable es de tipo numérica o categórica.

En el caso de variables categóricas se calcula el número de casos por cada categoría de la variable y el porcentaje que representa del total de registros. El objetivo es detectar elevadas concentraciones en determinadas categorías y, si existen casos no informados, qué porcentaje suponen estos respecto al total de registros de la base de datos.

Para las variables numéricas, se realiza un análisis descriptivo en el que se muestra un conjunto de estadísticas de resumen para evaluar las medidas de tendencia central. Las estadísticas son:

* Nmissing: Es el número de missings.
* Fill rate: Es el porcentaje de casos válidos.
* Mean: Es la media.
* Min: Es el mínimo valor de la variable.
* P25 (1st Qu.): Es el percentil 25.
* Median: Es la mediana o segundo cuartil.
* P75 (3erd Qu.): Es el percentil 75.



### Análisis temporal a nivel variable

Con este análisis se trata de detectar posibles cambios estructurales en la información. El análisis consiste en, para cada mes observado, analizar la distribución de la variable.

* Para las **variables numéricas** se calcula la media, mediana y los percentiles 25 y 75 los cuales se visualizan en un mismo gráfico para evaluar el comportamiento de cada estadística a través del tiempo (codmes).

* Para las **variables categóricas**, se calcula la distribución de la población en cada una de las categorías. Luego, mediante un gráfico de barras apiladas al 100% se evalúa si la concentración por categoría ha sufrido algún cambio significativo en algún punto del tiempo (codmes).



### Creación de Variables

Partiendo de las variables originales, se procede a definir nuevas variables de forma que la información a partir de la cual se califiquen las operaciones sea lo más enriquecedora posible.

El detalle de las variables creadas que finalmente se incluyen en el modelo se muestra en el documento de implementación adjunto:



Inicialmente partimos de 1,124 variables, posteriormente hacemos filtros de aquellas variablesque tienen missing< 0.05, nos quedamos con las 200 variables más importantes (gain) y sentido lógico de la variables versus target, finalmente quedándonos de la con un número reducido de variables de la siguiente forma:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | var iniciales | Missing <0.05 | Gain | sentido lógico | var usadas |
| N | 1,124 | 930 | 200 | 60 | 34 |

## ETAPA DE MODELAMIENTO

### Metodología de Modelamiento

### Gradient Boosting

Gradient Boosting utiliza un algoritmo de partición descrito en "Greedy Function Approximation: A Gradient Boosting Machine" y "Stochastic Gradient Boost" de Jerome H. Friedman. Un algoritmo de partición busca una partición óptima de los datos definidos en términos de los valores de una sola variable. El criterio de optimalidad depende de cómo otra variable, el objetivo, se distribuye en los segmentos de partición. Cuando los valores de la variable objetivo son más similares dentro de los segmentos, el valor de la partición es mayor. La mayoría de los algoritmos de partición particionan más cada segmento en un proceso llamado partición recursiva. Las particiones se combinan para crear un modelo predictivo. El modelo se evalúa mediante estadísticas de bondad de ajuste que se definen en términos de la variable objetivo. Estas estadísticas son diferentes de la medida del valor de una partición individual. Un buen modelo puede resultar de muchas particiones mediocres.

El Gradient Boosting es un enfoque de impulso que vuelve a muestrear el conjunto de datos de análisis varias veces para generar resultados que forman un promedio ponderado del conjunto de datos remuestreados. Los Arboles Boosting crea una serie de árboles de decisión que juntos forman un solo modelo predictivo. Un árbol de la serie se ajusta al residual de la predicción de los árboles anteriores de la serie. El residual se define en términos de la derivada de una función de pérdida.

Los objetivos de intervalo definen el residuo utilizando la función de pérdida de error al cuadrado. Para calcular el residual de un objetivo de intervalo utilizando la pérdida de error al cuadrado, simplemente reste el valor predicho del valor de destino. Los objetivos binarios definen el residuo utilizando la función de pérdida de probabilidad de registro binomial negativa. La función de pérdida de probabilidad binomial negativa también se conoce como pérdida logística.

Cada vez que se usan los datos para hacer crecer un árbol y se calcula la precisión del árbol. Las muestras sucesivas se ajustan para adaptarse a errores previamente calculadas. Debido a que cada muestra sucesiva se ponderó de acuerdo con la precisión de clasificación de los modelos anteriores, este enfoque a veces se denomina Gradient Boosting estocástico. El Boosting se define para objetivos binarios, nominales e intervalos.

Al igual que los árboles de decisión, El Boosting no hace suposiciones sobre la distribución de los datos. Para una entrada de intervalo, el modelo solo depende de los rangos de los valores. Para un objetivo de intervalo, la influencia de una teoría del valor extremo depende de la función de pérdida. El nodo Gradient Boosting ofrece una pérdida de la estimación de Huber M que reduce la influencia de los valores objetivo extremos. El Boosting es menos propenso a adaptar los datos que un solo árbol de decisiones, y si un árbol de decisiones se ajusta bastante bien a los datos, el aumento a menudo mejora el ajuste.

### Extreme Gradient Boosting (XGBoost)

XGBoost es una mejora del gradient boosting (GBM) a diferencia de su predecesor presenta un parámetro de regularización que controla el sobreajuste en el proceso aprendizaje automático.

Extreme Gradient Boosting o más conocido como XGBoost es un modelo de la familia de los Boosting basado en árboles de decisión o regresión según tipo de variable objetivo. El boosting es un tipo de método de ensamble que se basa en un proceso iterativo en el que en cada iteración del modelo minimiza el error predictivo y se va optiminizando hasta tener un pronosticador fuerte.

Además minimiza una función de objetivo regularizada (controla sobreajuste en cada iteración) que combina una función de pérdida convexa (según la diferencia entre las salidas de destino y previstas) y un plazo de penalización para la complejidad de modelos (es decir, las funciones de árboles de regresión). La capacitación avanza de forma iterativa, agregando nuevos árboles que predicen los residuos de errores de los árboles anteriores y se combinan después con los árboles anteriores para realizar la predicción final. Se denomina potenciación del gradiente porque utiliza un algoritmo de gradiente descendente para minimizar la pérdida cuando se agregan nuevos árboles de regresión. (Tianqi Chen - 2016).

La modelización consta de las siguientes etapas y fue desarrollado con el software Jupyter Hub - Python 3 en la nube.

### Partición de la muestra

Esta etapa se trabajó sin partición, es decir se trabajó el train al 100%.

Por otro lado, se trabajó con una muestra out of time con el objetivo de desarrollar modelos que puedan aplicarse adecuadamente a nuevos casos, distintos de los utilizados para el desarrollo. En efecto, si toda la muestra de análisis se utilizará para ajustar los modelos, se podría llegar a modelos que separasen muy bien los clientes buenos de los malos, pero no se tendría la garantía de que los resultados fueran similares a la hora de aplicar dichos modelos a nuevos casos. A continuación presentamos los totales de las particiones:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  | Train | OOT |  |
|  | N | 36,181 | 6,212 |  |
|  |  |  |  |  |

Finalmente se utiliza 13 cosechas de entrenamiento (202105, 202106, 202107, 202108, 202109, 202110, 202111, 202112, 202201, 202202, 202204, 202206, 202207) y 2 de validación fuera de tiempo (out of time) (202203,202205). Se usó los meses de Mar22 a May22 como oot, dado que en los últimos meses se observa una ligera aumento en la tasa, por ello se buscó tener un mejor modelo, incluyendo en el entrenamiento las tasas más altas y así obtener un modelo más estable y robusto en el tiempo.

Por otro lado resulta relevante mencionar que se realizó una imputación para los valores missing (ver descripción de variables) a las variables numéricas para poder diferenciarlas.

**Muestra de desarrollo:**

(202105,202106,202107,202108,202109,202110,202111,202112,202201,202202,202204,202206,202207)

|  |  |  |
| --- | --- | --- |
|  | **Total** | **%Total** |
| **Bueno** | 33,896 | 93.68% |
| **Malo** | 2,285 | 6.32% |
|  | 36,181 | 100% |

**Muestra de OOT**:

(202203,202205)

|  |  |  |
| --- | --- | --- |
|  | **Total** | **%Total** |
| **Bueno** | 5,851 | 94.19% |
| **Malo** | 361 | 5.81% |
|  | 6,212 | 100% |

### Modelo Hiperpámetros del Modelo Extreme Gradient Boosting

1. **Concepto de los Hiperparámetros**

* **Semilla (random\_state):** Especificar la semilla para generar las submuestras aleatorias en cada árbol.
* **Ratio Aprendizaje (learning\_rate)**: Use la propiedad Encogimiento para especificar cuánto reducir la predicción de cada árbol.
* **Profundidad Máxima (max\_depth)**: Máxima profundidad del árbol en cada iteración, cuanto más profundo tiende a sobreajustarse.
* **Proporción de entrenamiento (Bagging\_fraction/subsample)**: Proporción de submuestra de entrenamiento del árbol en cada iteración.
* **Proporción de variables (feature\_fraction/colsample\_bytree)**: Proporción de submuestra de variables al construir en cada árbol.
* **Gamma:** Reducción de pérdida mínima requerida para hacer una partición adicional en un nodo hoja del árbol.
* **N\_jobs:** Especifica número de subprocesos paralelos.
* **Tree Method:** Especifica el tipo de algoritmo a utilizar.
* **Objective:** Especifica el tipo del target.
* **Metric:** Especifica la métrica de perfomance a utilizar en el modelamiento.

1. **Cuadro de Hiperparámetros utilizados**:

Se entrenó el Algoritmo con los siguientes parámetros en el lenguaje de programación Python –Jupyter AWS:

|  |  |
| --- | --- |
| **Hiperparámetro** | **valor** |
| **colsample\_bylevel** | 1.0 |
| **gamma** | 1.5 |
| **learning\_rate** | 0.076 |
| **max\_delta\_step** | 0.5 |
| **reg\_alpha** | 0.7 |
| **reg\_lambda** | 1 |
| **max\_depth** | 2 |
| **min\_child\_weight** | 1.225 |
| **n\_estimators** | 230 |
| **subsample** | 0.51501 |
| **objective** | binarylogistic |
| **eval\_metric** | auc |
| **tree\_method** | auto |
| **n\_jobs** | -1 |
| **random\_state** | 123 |
| **booster** | gbtree |

1. **Grid Search** :

Consiste en ir probando con ensayo y error hasta la obtención del mejor hiperparámetro en cada intento de búsqueda a criterio experto.

Es una técnica manual que considera 3 aspectos:

* Mejor indicador de perfomance de predicción.
* Evitar sobreajuste.
* Estabilidad del modelo en el tiempo.

### Selección de variables

En primer lugar, se retiraron todas las variables que tuviesen menos de 0.5% de importancia relativizada por gain ( gain (i)/ Sumatoria (gain (n))), siendo i la ganancia por variable y n total de variables)

Luego de las variables restantes se hizo una selección por correlación, donde si una variable estaba fuertemente correlacionada con otra (>80%) se retiraba aquella que tuviera el menor % de importancia relativizada por shap value mean, el proceso es el siguiente:

1. Se reordenan las variables en cuestión por importancia relativizada gain
2. Se calcula la matriz de correlación de Pearson.
3. Se retiran todas las variables asociadas a una columna de la matriz de correlación donde el valor absoluto de la correlación sea mayor a determinado umbral (80%). Nótese que, como la base fue previamente ordenada y la matriz es triangular superior, si una pareja de variables sobrepasa el umbral de correlación aquella de menor importancia gain será la que esté en la columna de la matriz, es decir que se excluyen variables con alta correlación y presenten menor importancia.

Esto nos deja con un número reducido de variables, de todas formas podemos volver a las variables retiradas en este paso si posteriormente eliminamos alguna variable de esta lista final por criterio de negocio o experto.

Para llegar a las variables finales se utilizó la importancia por ganancia de modelo de xgboost tomando en cuenta el principio de parsimonia para escoger un número óptimo de variables en el modelo final.

### Análisis Bivariante/ Dependencia parcial con valores shap

El análisis bivariante se realiza con 3 objetivos:

* Estudiar la distribución de cada una de las variables candidatas a puntuar en el modelo, observando la presencia de valores extraños, el porcentaje de valores ausentes, las concentraciones en determinados valores, etc. Este análisis permite por un lado estimar la calidad de la información y por otro identificar variables que, por distintas causas, no reúnen los requisitos necesarios para ser consideradas en el análisis posterior.
* Obtener una primera aproximación del poder predictivo de la variable respecto a la variable indicador de incumplimiento a nivel univariado.
* Para las variables numéricas -al igual que para las variables categóricas- se calcula el número de casos, número de buenos, número de malos, la tasa de malos y el porcentaje de casos para cada uno de los grupos definidos.

Además de analizar las posibles concentraciones y el porcentaje de valores ausentes, también se estudia si la tendencia de la tasa de malos se corresponde con la esperada.

Por otro lado, tenemos que los SHapley Additive exPlanations(SHAP) son una metodología innovadora creada para poder darle explicabilidad tanto local (utilizando LIME) como global, apoyándose de la teoría de juegos, a todo tipo de modelo, unificando varios métodos previos de tal forma que representa el único método posible de atribución de características aditivas consistente y localmente preciso. SHAP asigna a cada variable una medida de importancia en cada predicción que depende de cuánto impacta en la predicción esperada del modelo condicionado a dicha variable. Aquí se presentan los valores SHAP asignados a cada variable para cada predicción en la muestra de validación de manera gráfica, lo cual nos da una vista sobre cómo es que el modelo trabaja con cada variable(Eje X) realmente, valores SHAP positivos (eje Y) implican un incremento en log-odds respecto al valor esperado de la predicción, en este caso particular representan un incremento en la probabilidad de default (superar más de 90 días de atraso en 1 año), mientras que valores SHAP negativos representan un decremento en dicha probabilidad.

En los siguientes archivos excels. Se detalla los gráficos bivariados de los shaps (los cuales fueron elaborados con el 100% de la data de TRAIN) y matriz de correlación de Pearson.

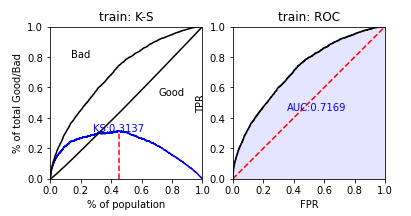


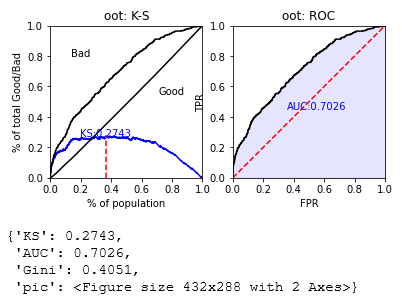
### Resultados de Perfomance del modelo

El gini del TRAIN fue calculado con el 100% de la base, y para el OOT se usó el periodo mencionado anteriormente, cabe mencionar que se establecen como criterios de rendimiento indicadores como Gini, KS y AUC.

Por otro lado, cabe resalta que se considera un modelo estable si la diferencia entre los indicadores de la muestra de desarrollo y oot es menor a 5 puntos porcentuales[[1]](#footnote-1)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **KS** | **AUC** | **GINI** |
| **TRAIN** | 31.37% | 71.69% | 43.39% |
| **OOT** | 27.43% | 70.26% | 40.51% |





**Importancia de variables por ganancia (gain)**

|  |  |  |  |
| --- | --- | --- | --- |
| Fuente | Variable | Descripción | Ganancia |
| Izipay | min\_trx\_3umv2 | Mínimo monto transaccionado en los 3um | 7.78% |
| Demográfica | edadv2 | Edad | 5.67% |
| Sunat | flag\_estado | Flag Activo | 5.18% |
| Izipay | rt\_1trx3umv2 | Ratio del monto transaccionado en el último mes respecto al promedio de monto transaccionado en los 3um | 5.03% |
| RCC RRLL | rt9sldnor12\_rlv2 | Ratio del promedio del saldo normal en los 9um con respecto al promedio del saldo normal en los 12um del rep legal | 4.41% |
| Sentinel | lvl\_eduv2 | Nivel educativo | 3.86% |
| Sentinel | prm\_ddatelc12mv2 | Identificador de promedio deuda en telecomunicaciones en los 12 meses | 3.59% |
| Sentinel | nro\_otremprep12mv2 | Número de deuda (cantidad ) de otras empresas en telefonía de los 12um | 3.53% |
| RCC consumo | rat\_sld\_cons\_3um\_6umv2 | Ratio del saldo consumo de los 3um con respecto al saldo consumo de los 6um | 3.43% |
| Sunarp | n\_meses\_anofabricacion\_nuevov2 | Antigüedad del vehículo más nuevo en meses | 3.38% |
| Adex | cnt\_impexp\_9umv2 | Cantidad de meses de importación y exportación en los 9um | 3.36% |
| RCC consumo | num\_disldcons\_9umv2 | numero de meses con disminución de saldo consumo en los 9um | 3.17% |
| Adex | min\_umvalor\_imp12umv2 | Mes más reciente del valor de importación de los 12um | 3.06% |
| RCC RRLL | max\_d\_sld\_6um\_rlv2 | Máxima disminución de saldo de los 6um del rep legal | 2.97% |
| RCC RRLL | num\_dsldnor\_12um\_rlv2 | Número de meses con disminución de saldo normal en los 12um del rep legal | 2.96% |
| RCC consumo | nument\_cons\_1umv2 | Número de entidades consumo en el último mes | 2.90% |
| RCC RRLL | rt9sldretail12\_rlv2 | Ratio del promedio del saldo consumo en los 9um con respecto al promedio del saldo consumo en los 12um del rep legal | 2.66% |
| Sentinel | prm\_telcorep12mv2 | Promedio de telecomunicaciones reportado de los últimos 12 meses | 2.52% |
| Adex | max\_a\_valorimp\_12umv2 | Máximo aumento del valor de importación de los 12um | 2.45% |
| Sunarp | cantidad\_veh | cantidad de vehículos | 2.43% |
| Sentinel | mnt\_ddaotremp12mv2 | Identificador de monto deuda de otra empresa 12 meses | 2.34% |
| RCC cónyuge | max\_num\_ent\_nor9um\_cnyv2 | Máximo número de entidades con clasificación normal en los 9um del cónyuge | 2.20% |
| Sunat | flag\_comercio | Flag comercio | 2.18% |
| RCC RRLL | min\_umdm12\_rlv2 | Mes más reciente con días de atraso de los 12um del rep legal | 2.03% |
| RCC consumo | max\_d\_sldcons\_9umv2 | Máxima disminución de saldo consumo en los 9um | 1.98% |
| RCC consumo | flag\_sobreg\_u3m | Flag sobregiro en los 3um | 1.92% |
| RCC cónyuge | max\_pc\_dsldnor\_12um\_cnyv2 | Máximo porcentaje de disminución de saldo normal en los 12um del cónyuge | 1.92% |
| RCC RRLL | max\_pc\_dsld\_6um\_rlv2 | Máximo porcentaje de disminución de saldo en los 6um del rep legal | 1.80% |
| Sunat | ant\_negociov2 | Antigüedad de negocio | 1.70% |
| RCC RRLL | rt3sldneg6\_rlv2 | Ratio del promedio de saldo negocio en los 3um en comparación con el promedio de saldo negocio en los 6um del rep legal | 1.64% |
| RCC consumo | num\_ausldcons\_6umv2 | Número de meses con aumento de saldo consumo en los 6um | 1.55% |
| RCC cónyuge | rt3sld6\_cnyv2 | Ratio del promedio del saldo total en los 3um en comparación con el promedio del saldo total en los 6um del cónyuge | 1.54% |
| RCC consumo | max12umcons\_dmv2 | Máximo días de atraso consumo de los 12um | 1.47% |
| Izipay | max\_autrx\_3umv2 | Máximo aumento del monto transaccionado en los 3um | 1.38% |

* 1. **Tabla de Eficiencias:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DECILES** | **BUENOS** | **MALOS** | **TOTAL** | **TM** | **%** | **TM ACUMULADA** |
| **(402.999, 863.0]** | 1294 | 466 | 1760 | 26% | 4% | 6.24% |
| **(863.0, 887.0]** | 1443 | 219 | 1662 | 13% | 4% | 5.37% |
| **(887.0, 900.0]** | 1563 | 213 | 1776 | 12% | 4% | 5.03% |
| **(900.0, 909.0]** | 1690 | 181 | 1871 | 10% | 4% | 4.70% |
| **(909.0, 916.0]** | 1546 | 131 | 1677 | 8% | 4% | 4.44% |
| **(916.0, 921.0]** | 1567 | 124 | 1691 | 7% | 4% | 4.27% |
| **(921.0, 926.0]** | 1600 | 131 | 1731 | 8% | 4% | 4.11% |
| **(926.0, 930.0]** | 1433 | 122 | 1555 | 8% | 4% | 3.91% |
| **(930.0, 934.0]** | 1548 | 104 | 1652 | 6% | 4% | 3.69% |
| **(934.0, 938.0]** | 1614 | 119 | 1733 | 7% | 4% | 3.53% |
| **(938.0, 942.0]** | 1646 | 125 | 1771 | 7% | 4% | 3.31% |
| **(942.0, 946.0]** | 1883 | 113 | 1996 | 6% | 5% | 3.02% |
| **(946.0, 949.0]** | 1611 | 65 | 1676 | 4% | 4% | 2.78% |
| **(949.0, 952.0]** | 1689 | 83 | 1772 | 5% | 4% | 2.69% |
| **(952.0, 954.0]** | 1222 | 50 | 1272 | 4% | 3% | 2.49% |
| **(954.0, 957.0]** | 1848 | 82 | 1930 | 4% | 5% | 2.38% |
| **(957.0, 959.0]** | 1310 | 47 | 1357 | 3% | 3% | 2.14% |
| **(959.0, 962.0]** | 1885 | 61 | 1946 | 3% | 5% | 2.01% |
| **(962.0, 965.0]** | 1962 | 48 | 2010 | 2% | 5% | 1.82% |
| **(965.0, 967.0]** | 1347 | 35 | 1382 | 3% | 3% | 1.70% |
| **(967.0, 970.0]** | 1767 | 38 | 1805 | 2% | 4% | 1.55% |
| **(970.0, 973.0]** | 1763 | 34 | 1797 | 2% | 4% | 1.40% |
| **(973.0, 976.0]** | 1366 | 27 | 1393 | 2% | 3% | 1.20% |
| **(976.0, 981.0]** | 1765 | 23 | 1788 | 1% | 4% | 0.88% |
| **(981.0, 996.0]** | 1385 | 5 | 1390 | 0% | 3% | 0.36% |

### Sofware utilizado:

A continuación se presenta el software utilizado para todos los procesos Mencionados:

1. Creación de bases: Teradata SQL Assisttant version 15
2. Bases en la nube: Archivo parquet S3 – AWS
3. Modelamiento: Jupyter Hub - Python 3

## ANEXOS

* Anexo 1

|  |  |
| --- | --- |
|  | Modelo |
| Target |  |
| Diccionario de variables |  |
| Objeto del Modelo XGBOOST |  |

1. Dentro de los márgenes que manejamos el banco aceptamos una caída máxima de 5 puntos. [↑](#footnote-ref-1)