**MODELO COMPORTAMENTAL BPE**

Área Modelos y Metodologías

SQUAD BREAKING SCHEMES

**INTERBANK**

**Control de Versiones**

|  |  |  |  |
| --- | --- | --- | --- |
| **Versión** | **Fecha** | **Realizado por** | **Actualización realizada** |
| 1 | 10-10-2022 | Angel Rivera | Diseño del modelo |
| 2 | 19-10-2022 | Angel Rivera | Modelos de los segmento reactiva y reprogramados |
| 3 | 21-10-2022 | Angel Rivera | Documentación con los 3 modelos |

Contenido

[1. INTRODUCCIÓN 3](#_Toc117249335)

[2. POBLACIÓN OBJETIVO 3](#_Toc117249336)

[3. RECOPILACIÓN Y GENERACIÓN DE INFORMACIÓN 5](#_Toc117249337)

[3.1 Fuentes de información requeridas 5](#_Toc117249338)

[3.2 Softwares utilizados 7](#_Toc117249339)

[4. DESCRIPCION DE PROCESOS 7](#_Toc117249340)

[4.1 OBTENCION DE LA MUESTRA 7](#_Toc117249341)

[4.2 FITLTROS APLICADOS SOBRE LA BASES DE DATOS INICIALES 7](#_Toc117249342)

[4.3 PERIODOS DE OBSERVACIÓN 8](#_Toc117249343)

[4.2 SEGMENTACIÓN 9](#_Toc117249344)

[4.3 PERIODO DE COMPORTAMIENTO 10](#_Toc117249345)

[5. MARCA DE DEFAULT 11](#_Toc117249346)

[5.1 Análisis de segmentación: 12](#_Toc117249347)

[5.2 Análisis descriptivo 13](#_Toc117249348)

[5.3 Análisis temporal a nivel variable 14](#_Toc117249349)

[5.4 Creación de Variables 15](#_Toc117249350)

[6. ETAPA DE MODELAMIENTO 16](#_Toc117249351)

[6.1 Metodología de Modelamiento 16](#_Toc117249352)

[Gradient Boosting 16](#_Toc117249353)

[Extreme Gradient Boosting (XGBoost) 17](#_Toc117249354)

[6.2 Partición de la muestra 18](#_Toc117249355)

[6.3 Modelo Hiperpámetros del Modelo Extreme Gradient Boosting 20](#_Toc117249356)

[6.4 Selección de variables 23](#_Toc117249357)

[6.5 Análisis Bivariante/ Dependencia parcial con valores shap 24](#_Toc117249358)

[6.6 Resultados de Perfomance del modelo 25](#_Toc117249359)

[6.8 Sofware utilizado: 36](#_Toc117249360)

[7. ANEXO 36](#_Toc117249361)

[8. RESPONSABILIDAD DE ÁREA Y ÓRGANOS INTERVENIENTES 37](#_Toc117249362)

## INTRODUCCIÓN

El modelo que actualmente evalúa a los clientes de la cartea de Banca Pequeña Empresa han presentado un deterioro en el poder predictivo, reduciendo hasta 24ppc el indicador del Gini, debido principalmente al impacto económico producido por la pandemia y las acciones tomadas por el gobierno a través de préstamos reactiva, reprogramaciones, congelamientos, etc. Es por eso, que surge la necesidad de desarrollar un nuevo modelo que cumpla nuevamente con los estándares y niveles de predicción esperados para este portafolio (ver anexo 7).

El nuevo Score comportamental BPE es un Modelo matemático que clasifica a un cliente según su perfil de riesgo, en alguna de las siguientes categorías: bueno o malo.

La clasificación del score se basa en la probabilidad de caer en default (incumplimiento). Esta probabilidad se representa como un número, que es calculado como suma de los coeficientes estimados para cada una de las variables que perfilan al cliente; estos coeficientes son transformados en puntajes obteniendo como resultado un score que valora el riesgo del cliente.

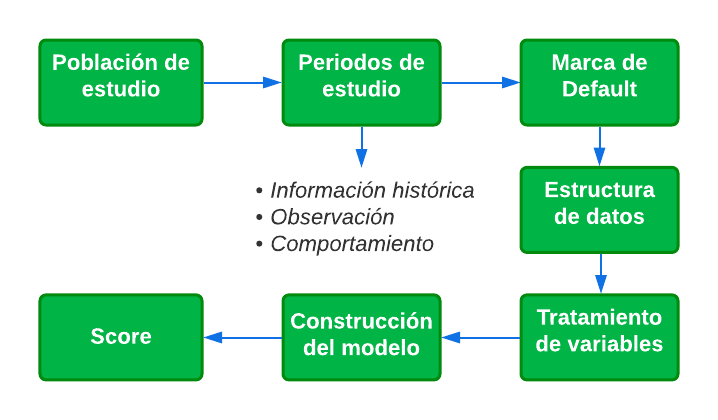
## POBLACIÓN OBJETIVO

La población objetivo son todos los clientes BPE (definimos como BPE a aquellas empresas con niveles de ventas entre 180 000 y 3 millones soles anuales, con posibles excepciones) con deuda BPE vigente en Interbank con algún producto (Línea revolvente, Capital de trabajo, etc.), aplicando las siguientes consideraciones:

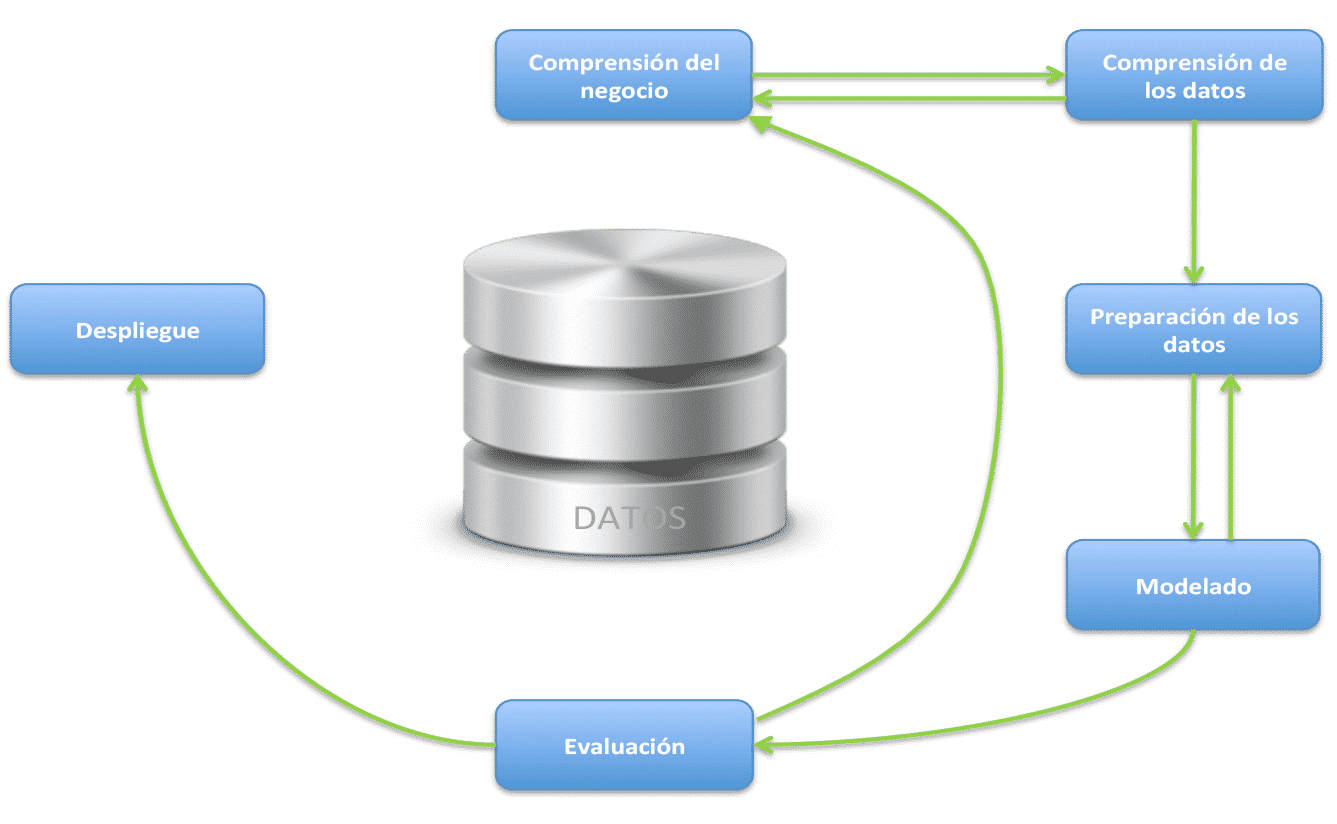
* Clientes con una antigüedad mínima de 3 meses en Interbank.

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, prejudicial, judiciales, castigados en el periodo de observación.
* Clientes con días mora en el periodo de observación mayor a 30 días.



Metodología CRIPS-DM



1. **Comprensión del Negocio**: Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto.
2. **Comprensión de los datos**: Fase de entendimiento de los datos donde realizamos distintos análisis para poder identificar los periodos y el universo de desarrollo del modelo.
3. **Preparación de los datos**: Fase que cubre todas las actividades de generación de variables de las distintas fuentes que manejamos.
4. **Modelado**: En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos.
5. **Evaluación**: Fase en donde se evalúa el modelo en distintos escenarios, se revisa a profundidad el sentido de negocio de cada una de las variables y se trabaja en las estrategias del modelo.
6. **Despliegue**: Fase donde el modelo se pone en producción.

## 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.



RCC

Sunat

Adex

BPE- CF





Trx.

Emp.

Pasivos

Reniec

Cuotas



Provee

dores

* **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 Cambio de Fecha – Cuotas:** Bases internas de Interbank con información de solicitudes de cambio de fecha para el pago de una cuota y cronograma de pagos respectivamente.
* **Base Reniec:** Variables que corresponden a la información registrada en fuentes internas del banco donde se identifica información de tipo sociodemográfica e identidad del cliente. Base comprada.
* **Base Pasivos (Base Interna):** Variables que involucran cuentas de pasivos por parte del cliente, donde se puede identificar cuentas de pasivo en Interbank.
* **Base Transacciones Empresa (Base Interna):** Base donde se encuentran todas las transacciones de las cuentas de ahorro que salen e ingresan de las empresas.
* **Base Proveedores (Base Interna):** Base que contiene la información de pagos realizados a los proveedores (potenciales clientes BPE) de nuestros clientes de la Banca Empresa o Corporativa.

**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\_dw\_views** | v\_mst\_vpc\_adex | MENSUAL |  |
|  | **e\_dw\_views** | v\_fact\_vpc\_ppa\_transacc\_inof | DIARIO |  |
|  | **e\_dw\_views** | v\_agg\_mes\_vpc\_transacc\_sav\_imp | DIARIO |  |
|  | **e\_perm\_aws** | t\_fem\_all\_instr\_rsk | MENSUAL |  |
|  | **e\_dw\_views** | v\_agg\_cambiofecha\_hist\_rsk | MENSUAL |  |
|  | **e\_perm\_aws** | t\_agg\_cuotmype\_hist\_rsk t\_agg\_cuotleasing\_hist\_rsk | MENSUAL |  |
|  | **e\_perm\_aws** | t\_mst\_reniec\_rsk | ANUAL |  |
|  |  |  |  |  |

### 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 contrucción del modelo donde se usaron paquetes como pandas (versión 1.3.4), numpy (versión 1.20.3), xgboost (versión 1.4.2), scorecardpy (versión 0.1.9.2).

## DESCRIPCION DE PROCESOS

### OBTENCION DE LA MUESTRA

Con el objetivo de tener un modelo que capte el comportamiento de los nuevos perfiles surgidos por la coyuntura covid, se van a considerar solo periodos del 2021 como periodos de comportamiento. Los períodos mostrados en los siguientes gráficos hacen referencia a la fecha de cierre.

### FITLTROS APLICADOS SOBRE LA BASES DE DATOS INICIALES

* *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, deben de tener valor 0 en los campos prejudicial, judicial, castigado y refinanciado en Interbank.

* *Indeterminados en el origen (flag\_ind\_orig=1):*

**Motivo**. Son clientes que son indeterminados al momento de la observación por lo que aquellos clientes que posean mayor a 30 días en Interbank son excluidos de la población objetivo.

* *Sin deuda (flag\_deuda=1):*

**Motivo**. Clientes que no presentan deuda en el periodo de observación.

* *Antigüedad del cliente (flag\_ant=1):*

**Motivo**: Son clientes que tienen como mínimo 3 meses de antigüedad en Interbank.

* *Meses con performance sin congelar (flag\_meses\_pef=1):*

**Motivo**: Algunos créditos de BPE cuentan con periodos de gracia debido a eso nos estamos quedando con aquellos clientes que presentan como mínimo 6 meses de performance sin ser afectados por los periodos de gracia.

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

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

* *Indeterminados (flag\_malo=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 los clientes que alcanzan más de 30 días, pero no los 91 días.

|  |  |  |
| --- | --- | --- |
| **Exclusiones y filtros Target** | **Periodo** |  |
| **202103 - 202106** | **Filtrados** |
| Universo BPE | 132,058 |  |
| Malos en el origen | 114,126 | 17,932 |
| Ind en el origen | 108,493 | 5,633 |
| Sin deuda | 108,373 | 120 |
| Antigüedad del cliente | 103,144 | 5,229 |
| Meses de performance | 101,004 | 2,140 |
| Insuficientes | 96,641 | 4,363 |
| Indeterminados | 76,569 | 20,072 |
| Base final | 76,569 | - |

### PERIODOS DE OBSERVACIÓN

La muestra se desarrolló con una ventana de observación conformada por 4 cosechas tomando como periodos desde Mar-2021 hasta Jun-2021, no se consideraron los periodos de Ene-2021 y Feb-2021 debido a que el 60% de los clientes presentaban como mínimo 7 meses de su performance afectado por los periodos gracia o por el congelamiento de reactiva (ver anexo 7) y eso generaban una inestabilidad en volumen y en target. A continuación una vista global:

|  |  |  |
| --- | --- | --- |
| codmes | Total | TM |
| 202103 | 18257 | 22.5% |
| 202104 | 19302 | 22.6% |
| 202105 | 19573 | 23.4% |
| 202106 | 19437 | 23.5% |

### SEGMENTACIÓN

Se realizaron 3 segmentaciones para una gestión más adecuada, estos son:

1. **Segmento Negocio + Mixto:** Universo conformado por aquellos clientes que presentan solo créditos BPE o que presentan un crédito reactiva[[1]](#footnote-1) aparte de su crédito BPE (Mixto). Los créditos reactiva que ingresan en este segmento no presentan reprogramaciones.
2. **Segmento Reactiva:** Universo conformado solamente por aquellos clientes que cuentan solo con créditos reactiva. De la misma manera que en el anterior segmento, las reactivas que ingresan en este segmento no están reprogramados.
3. **Segmento Reprogramado:** Universo conformado por todos los clientes que presentan un crédito reactiva y que además hallan reprogramado dicho crédito.

A continuación una vista de cada uno de los segmentos:

**Segmento Negocio + Mixto:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CODMES** | **0** | **1** | Total | TM |
| **202103** | 4274 | 1327 | 5601 | 23.7% |
| **202104** | 4364 | 1327 | 5691 | 23.3% |
| **202105** | 4443 | 1363 | 5806 | 23.5% |
| **202106** | 4434 | 1232 | 5666 | 21.7% |
| **Total** | 17515 | 5249 | 22764 | 23.1% |

**Segmento Reactiva:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CODMES** | **0** | **1** | Total | TM |
| **202103** | 3825 | 1621 | 5446 | 29.8% |
| **202104** | 4120 | 1779 | 5899 | 30.2% |
| **202105** | 4096 | 1859 | 5955 | 31.2% |
| **202106** | 4037 | 1918 | 5955 | 32.2% |
| **Total** | 16078 | 7177 | 23255 | 30.9% |

**Segmento Reprogramado:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CODMES** | **0** | **1** | Total | TM |
| **202103** | 6055 | 1155 | 7210 | 16.0% |
| **202104** | 6461 | 1251 | 7712 | 16.2% |
| **202105** | 6461 | 1351 | 7812 | 17.3% |
| **202106** | 6406 | 1410 | 7816 | 18.0% |
| **Total** | 25383 | 5167 | 30550 | 16.9% |

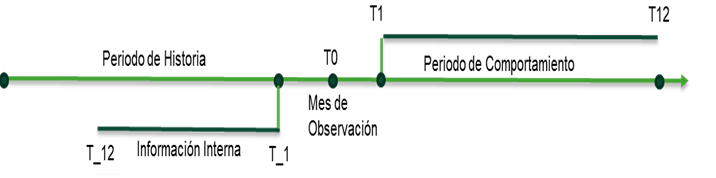
### PERIODO DE COMPORTAMIENTO

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 12 meses siguientes respecto a cada mes de observación, que comprenden los periodos de Mar-2022 hasta Jun-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, aquella persona (natural o jurídica) que presenta alguna de las siguientes condiciones:

* Si posee una o más créditos refinanciadas
* Si posee una o más créditos castigados
* Si está en los tramos de prejudicial y judicial.
* Si presenta atraso mayor a 90 días (>90) en al menos 1 mes en los 12 siguientes meses (ventana de observación)

Estos días de mora, refinanciando y castigados considera todos los productos de BPE a nivel Interbank.

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 de segmentación:

Debido a la naturaleza de los clientes, es necesario establecer modelos diferentes para los 3 segmentos (Negocio + Mixto, Reactiva, Reprogramados), esta diferencia se puede ver reflejada tanto en la definición de universos como en la distribución de la Tasa de Malos por periodo.

|  |  |  |
| --- | --- | --- |
| Segmento | Total | Porcentaje |
| Neg + Mixto | 22,764 | 30% |
| Reactiva | 23,255 | 30% |
| Reprogramados | 30,550 | 40% |

### 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.

**Segmento Negocio + Mixto**



**Segmento Reactiva**



**Segmento Reprogramado**



### 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).

**Segmento Negocio + Mixto**



**Segmento Reactiva**



**Segmento Reprogramado**

****

### 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:

**Segmento Negocio + Mixto**



**Segmento Reactiva**



**Segmento Reprogramado**



Inicialmente partimos de 3,660 variables, retiramos aquellas variables que han sido construidas en un mes específico (variables preliminares), luego hacemos filtros (por segmentos) de aquellas variables que tienen un porcentaje mayor al 95% de missings. De éstas variables se seleccionan mediante criterios explicados en la sección 6.4.

**Segmento Negocio + Mixto:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | var iniciales | Var preliminares | missing |
| N | 3,666 | 1,727 | 1,526 |

**Segmento Reactiva:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | var iniciales | Var. preliminar | missing |
| N | 3,666 | 1,727 | 1,399 |

**Segmento Reprogramado:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | var iniciales | Var. preliminar | missing |
| N | 3,666 | 1,727 | 1,398 |

## 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 con una participación 80-20 (train-test).

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:

**Segmento Negocio + Mixto:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  | Train | Test | OOT |  |
|  | N | 13,566 | 3,392 | 5,806 |  |
|  |  |  |  |  |  |

**Segmento Reactiva:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  | Train | Test | OOT |  |
|  | N | 13,840 | 3,460 | 5,955 |  |

**Segmento Reprogramado:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  | Train | Test | OOT |  |
|  | N | 18,190 | 4,548 | 7,8012 |  |

Finalmente se utiliza 3 cosechas de entrenamiento y 1 de validación fuera de tiempo (out of time).

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 y observar que el modelo los esté castigando.

**Segmento Negocio + Mixto:**

**Muestra de desarrollo** (202103-202104,202106):

|  |  |  |
| --- | --- | --- |
|  | **Total** | **%Total** |
| **Bueno** | **13072** | **77.1%** |
| **Malo** | **3886** | **22.9%** |
|  | **16958** | **100%** |

**Muestra de OOT** (202105):

|  |  |  |
| --- | --- | --- |
|  | **Total** | **%Total** |
| **Bueno** | **4443** | **76.5%** |
| **Malo** | **1363** | **23.5%** |
|  | **5806** | **100%** |

**Segmento Reactiva:**

**Muestra de desarrollo** (202103-202104,202106):

|  |  |  |
| --- | --- | --- |
|  | **Total** | **%Total** |
| **Bueno** | **11982** | **69.3%** |
| **Malo** | **5318** | **30.7%** |
|  | **17300** | **100%** |

**Muestra de OOT** (202105):

|  |  |  |
| --- | --- | --- |
|  | **Total** | **%Total** |
| **Bueno** | **4096** | **68.8%** |
| **Malo** | **1859** | **31.2%** |
|  | **5955** | **100%** |

**Segmento Reprogramado:**

**Muestra de desarrollo** (202103-202104,202106):

|  |  |  |
| --- | --- | --- |
|  | **Total** | **%Total** |
| **Bueno** | **18922** | **83.2%** |
| **Malo** | **3816** | **16.8%** |
|  | **22738** | **100.0%** |

**Muestra de OOT** (202105):

|  |  |  |
| --- | --- | --- |
|  | **Total** | **%Total** |
| **Bueno** | **6461** | **82.7%** |
| **Malo** | **1351** | **17.3%** |
|  | **7812** | **100.0%** |

### Modelo Hiperpámetros del Modelo Extreme Gradient Boosting

1. **Concepto de los Hiperparámetros**

* **Numero Iteraciones (num\_boost\_round):** Especificar el número de iteraciones en la serie de Boosting. Para objetivos de intervalo y binarios, el número de iteraciones es igual a los árboles. Para un objetivo nominal, se crea un árbol separado para cada categoría de destino en cada serie de iteraciones.
* **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.
* **Early\_stopping\_rounds:** Especifica el número de veces hasta encontrar el mejor perfomance entre train-test.
* **monotone\_constraints:** Nos permite definir previamente que relación monótona va tener la variable con respecto al target 0 sin restricción monótona, 1: restricción monótona creciente, -1: restricción monótona decreciente.

1. **Cuadro de Hiperparámetros utilizados**:

Se entrenó el Algoritmo con los siguientes parámetros en el lenguaje de programación Python 3 – Jupyter Lab:

**Segmento Negocio + Mixto:**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | **Hiperparámetro** | **valor** |  |
|  | **colsample\_bylevel** | 0.95 |  |
|  | **colsample\_bytree** | 0.75 |  |
|  | **monotone\_constraints** | (-1,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,-1,0)[[2]](#footnote-2) |  |
|  | **gamma** | 0.95 |  |
|  | **learning\_rate** | 0.15 |  |
|  | **max\_delta\_step** | 1 |  |
|  | **reg\_lambda** | 1 |  |
|  | **max\_depth** | 3 |  |
|  | **min\_child\_weight** | 2 |  |
|  | **n\_estimators** | 369 |  |
|  | **subsample** | 0.99 |  |
|  | **n\_jobs** | -1 |  |
|  | **objective** | binarylogistic |  |
|  | **eval\_metric** | auc |  |
|  | **silent** | 1 |  |
|  | **tree\_method** | auto |  |
|  | **seed** | 33 |  |
|  | **early\_stopping\_rounds** | 5 |  |
|  | **booster** | gbtree |  |
|  | **Mejor iteración** | 115 |  |

**Segmento Reactiva:**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | **Hiperparámetro** | **valor** |  |
|  | **colsample\_bylevel** | 0.95 |  |
|  | **colsample\_bytree** | 0.75 |  |
|  | **monotone\_constraints** | (-1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)[[3]](#footnote-3) |  |
|  | **gamma** | 0.95 |  |
|  | **learning\_rate** | 0.16 |  |
|  | **max\_delta\_step** | 1 |  |
|  | **reg\_lambda** | 1 |  |
|  | **max\_depth** | 3 |  |
|  | **min\_child\_weight** | 2 |  |
|  | **n\_estimators** | 369 |  |
|  | **subsample** | 0.71 |  |
|  | **n\_jobs** | -1 |  |
|  | **objective** | binarylogistic |  |
|  | **eval\_metric** | auc |  |
|  | **silent** | 1 |  |
|  | **tree\_method** | auto |  |
|  | **seed** | 33 |  |
|  | **early\_stopping\_rounds** | 5 |  |
|  | **booster** | gbtree |  |
|  | **Mejor iteración** | 131 |  |

**Segmento Reprogramado:**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | **Hiperparámetro** | **valor** |  |
|  | **colsample\_bylevel** | 0.98 |  |
|  | **colsample\_bytree** | 0.75 |  |
|  | **monotone\_constraints** | (0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0,0,0,0,0,0)[[4]](#footnote-4) |  |
|  | **gamma** | 0.95 |  |
|  | **learning\_rate** | 0.14 |  |
|  | **max\_delta\_step** | 1 |  |
|  | **reg\_lambda** | 1 |  |
|  | **max\_depth** | 3 |  |
|  | **min\_child\_weight** | 2 |  |
|  | **n\_estimators** | 369 |  |
|  | **subsample** | 0.79 |  |
|  | **n\_jobs** | -1 |  |
|  | **objective** | binarylogistic |  |
|  | **eval\_metric** | auc |  |
|  | **silent** | 1 |  |
|  | **tree\_method** | auto |  |
|  | **seed** | 33 |  |
|  | **early\_stopping\_rounds** | 5 |  |
|  | **booster** | gbtree |  |
|  | **Mejor iteración** | 194 |  |

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

El total de variables está conformado por distintas fuentes mencionadas al inicio del documento. A medida que se tenían las variables de cada fuente, se seguía el siguiente proceso a nivel de segmentos:

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.

**Segmento Negocio + Mixto:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | var iniciales | Var preliminares | missing | correlación | gain | Var finales |
| N | 3,660 | 1,727 | 1,526 | 800 | 400 | 20 |

**Segmento Reactiva:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | var iniciales | Var. preliminar | missing | correlación | gain | Var finales |
| N | 3,666 | 1,727 | 1,399 | 700 | 350 | 24 |

**Segmento Reprogramado:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | var iniciales | Var. preliminar | missing | correlación | gain | Var finales |
| N | 3,666 | 1,727 | 1,398 | 700 | 350 | 19 |

### 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.

**Segmento Negocio + Mixto:**



**Segmento Reactiva:**



**Segmento Reprogramado:**



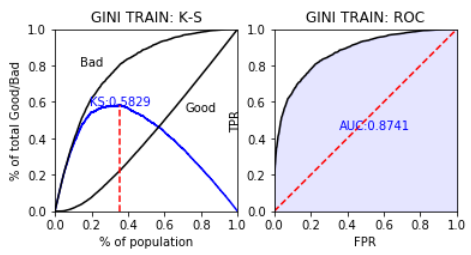
### Resultados de Perfomance del modelo

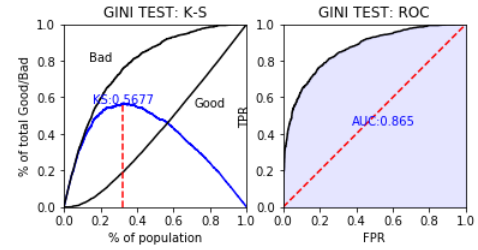
El gini del TRAIN y TEST fue calculado con el 80% y 20% de la base respectivamente para cada segmento, y para el OOT se usaron los periodos mencionados anteriormente, cabe mencionar que se establecen como criterios de rendimiento indicadores como Gini, KS y AUC.

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.

**Segmento Negocio + Mixto:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  | **KS** | **AUC** | **GINI** |  |
|  | **TRAIN** | 58.3% | 87.4% | 74.8% |  |
|  | **TEST** | 56.8% | 86.5% | 73.0% |  |
|  | **OOT** | 56.9% | 86.8% | 73.7% |  |
|  |  |  |  |  |  |

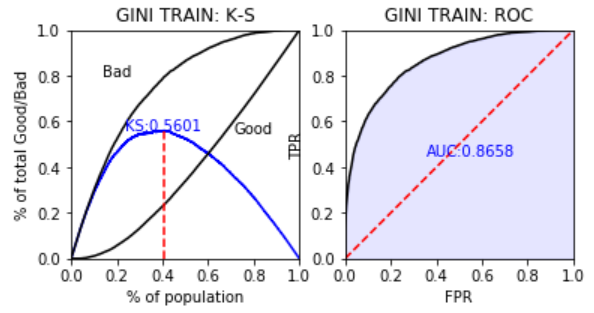


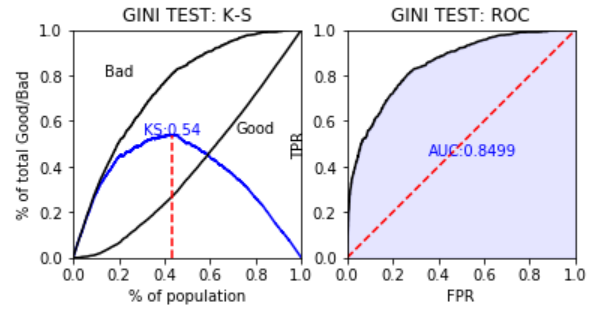


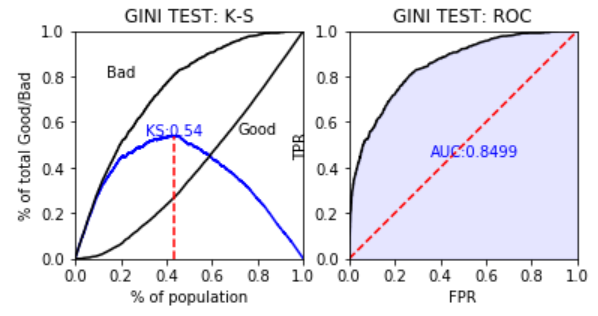


**Segmento Reactiva:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  | **KS** | **AUC** | **GINI** |  |
|  | **TRAIN** | 56.01% | 86.6% | 73.2% |  |
|  | **TEST** | 54.00% | 85.0% | 70.0% |  |
|  | **OOT** | 54.52% | 85.46% | 70.9% |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

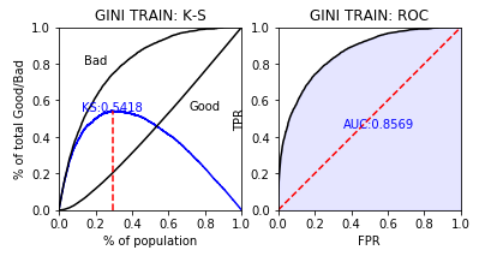


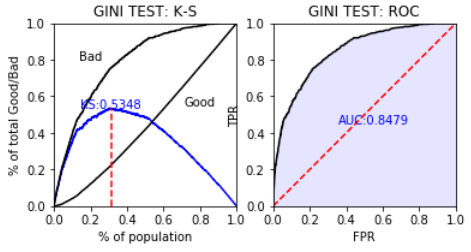


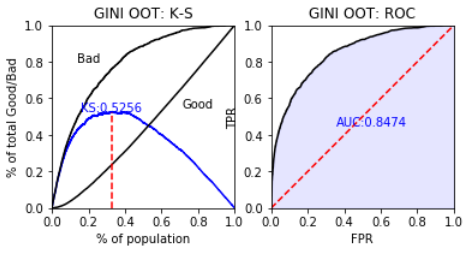


**Segmento Reprogramado:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  | **KS** | **AUC** | **GINI** |  |
|  | **TRAIN** | 54.18% | 85.7% | 71.4% |  |
|  | **TEST** | 53.48% | 84.8% | 69.6% |  |
|  | **OOT** | 52.56% | 84.7% | 69.5% |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |







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

**Segmento Negocio + Mixto:**

|  |  |  |
| --- | --- | --- |
| **Descripción** | **Ganancia** | **Fuentes** |
| Promedio de la participación del saldo no normal en los 6um | 12% | RCC |
| Máximo días de atraso intrames 6um | 11% | IBK |
| Máximo número de entidades en clasificación no normal | 9% | RCC |
| Mes mas reciente que presento atraso en créditos negocio en los 6um | 8% | RCC |
| Número de meses con mora en los 6um | 7% | RCC |
| Indicador si el cliente solicito un cambio de fecha en los 9um | 5% | IBK |
| Número de meses que disminuye el saldo en crédito negocio en los 6um | 5% | RCC |
| Máximo aumento en porcentaje del saldo no vigente en los 3um | 5% | RCC |
| Peor clasificación del Rep. Legal 1um | 4% | RCC - SUNAT |
| Indicador si la empresa esta activo | 4% | SUNAT |
| Agrupamiento del ratio entre prm pasivo 6um / prm pasivo 12um | 3% | PASIVOS |
| Antigeudad en IBK | 3% | IBK |
| Antigüedad del ruc + Promedio del saldo pasivo 6um | 3% | SUNAT - PASIVOS |
| Número de meses que realiza exp o imp en los 9um | 3% | ADEX |
| Máximo porcentaje de disminución del saldo en los 6um | 3% | RCC |
| Máximo días de atraso en los 3um del rep. Legal | 3% | RCC - SUNAT |
| Número de meses que adelanto pago de cuota en los 9um | 3% | IBK |
| Número de meses con mora respecto al conyugue en los 9um | 2% | RCC - RENIEC |
| Ratio entre el prm del pasivo prm 6um / prm del pasivo prm 12um | 2% | PASIVOS |
| Máximo número de entidades no normal del conyugue en los 9um | 2% | RCC - RENIEC |
| Agrupamiento del promedio del cash in promedio de los 6um | 2% | TRX. EMPRESA |
| Mínimo saldo pasivo en los 12um | 1% | PASIVOS |

**Segmento Reactiva:**

|  |  |  |
| --- | --- | --- |
| **Descripción** | **Ganancia** | **Fuentes** |
| Indicador si la empresa está activo | 16% | SUNAT |
| Máximo número de entidades no normal en los 6um Rep. Legal | 14% | RCC-SUNAT |
| Indicador si el Rep. Legal presenta saldo castigado | 10% | RCC-SUNAT |
| Número de meses con pasivos mayores a 10 soles en los 6um | 7% | PASIVOS |
| Mínimo saldo pasivo promedio en los 3um | 6% | PASIVOS |
| Ratio entre el prm pasivo prm 6um / el prm pasivo prm 12 | 5% | PASIVOS |
| Mes más reciente que presento atraso el Rep. Legal en los 6um | 5% | RCC-SUNAT |
| Agrupamiento del máximo días de atraso del Rep. Legal en los 12um | 4% | RCC-SUNAT |
| Número de meses que presenta mora en créditos retail el Rep. Legal 12um | 4% | RCC-SUNAT |
| Antigüedad del negocio | 4% | SUNAT |
| Número de meses que incrementa los pasivos en 100 soles en los 9um | 4% | PASIVOS |
| Indicador si realiza importación y exportación en los 6um | 3% | ADEX |
| Indicador si la empresa realiza pago de planilla en los 6um | 3% | TRX. EMPRESA |
| Promedio del porcentaje de cash out en trasferencia en los 12um | 3% | TRX. EMPRESA |
| Número de meses con mora del conyugue en los 6um | 2% | RCC-RENIEC |
| Promedio del % de cash out en los 3um | 2% | TRX. EMPRESA |
| Máximo disminución del porcentaje del saldo retail del Rep. Legal | 2% | RCC-SUNAT |
| Número de meses que disminuye el saldo del Rep. Legal en los 12um | 2% | RCC-SUNAT |
| Número de meses que realizas un cash out por pagos de servicios en los 6um | 1% | TRX. EMPRESA |
| Número de meses que incrementa el ingreso de los proveedores en los 6um | 1% | PROVEEDORES |
| Ratio entre el prm de cash in 3um/ prm del cash in 6um | 1% | TRX. EMPRESA |
| Número de meses que incrementa el ingreso del cash in por transferencia en los 6um | 1% | TRX. EMPRESA |

**Segmento Reprogramado:**

|  |  |  |
| --- | --- | --- |
| **Descripción** | **Ganancia** | **Fuentes** |
| Número de meses con pasivos mayores a 10 soles en los 6um | 14% | PASIVOS |
| Peor clasificación en los 3um del Rep. Legal | 12% | RCC - SUNAT |
| Indicador si la empresa está activa | 11% | SUNAT |
| Ratio entre el promedio 6um / promedio 12um | 7% | PASIVOS |
| Indicador si el rep. Legal presenta deuda castigada en los 9um | 6% | RCC-SUNAT |
| Mes más reciente que presenta atraso en créditos negocio en los 6um | 6% | RCC |
| Indicador si presenta entidad en clasificación no normal en los 3um | 5% | RCC |
| Número de meses que incrementa sus pasivos como mínimo 100 soles en los 6um | 4% | PASIVOS |
| Promedio de pasivos en los 6um | 4% | PASIVOS |
| Mes más reciente que presenta atraso en los 6um el Rep. Legal | 4% | RCC-SUNAT |
| Máximo días de atraso con respecto al conyugue en los 6um | 3% | RCC-RENIEC |
| Máximo número de entidades no normal en los 3um del Rep. Legal | 3% | RCC-SUNAT |
| Agrupamiento del mínimo saldo promedio pasivo en los 12um | 3% | PASIVOS |
| Antigüedad del ruc | 3% | SUNAT |
| Promedio del % del saldo normal relacionado al conyugue en los 6um | 3% | RCC-RENIEC |
| Número de meses que incrementa los ingresos de los proveedores en los 6um | 2% | PROVEEDORES |
| Promedio del porcentaje del cash out por transferencia en los 12um | 2% | TRX. EMPRESA |
| Promedio del % de cash out en los 3um | 2% | TRX. EMPRESA |
| Agrupamiento del ratio en el promedio del cash in 3um/ promedio del cash in 6um | 2% | TRX. EMPRESA |
| Indicador si la empresa realiza importación o exportación en los 6um | 2% | ADEX |
| Número de meses que disminuye el saldo en los 6um del Rep. Legal | 2% | RCC-SUNAT |

* 1. **Tabla de Eficiencias:**

**Segmento Negocio + Mixto**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DECILES** | **BUENOS** | **MALOS** | **TOTAL** | **TM** | **%** | **TM ACUMULADA** |
| **(0.999, 136.0]** | 21 | 892 | 913 | 98% | 4% | 23.06% |
| **(136.0, 312.0]** | 153 | 763 | 916 | 83% | 4% | 19.94% |
| **(312.0, 448.56]** | 297 | 606 | 903 | 67% | 4% | 17.17% |
| **(448.56, 557.0]** | 380 | 537 | 917 | 59% | 4% | 14.92% |
| **(557.0, 633.0]** | 506 | 405 | 911 | 44% | 4% | 12.82% |
| **(633.0, 689.0]** | 612 | 301 | 913 | 33% | 4% | 11.24% |
| **(689.0, 731.0]** | 660 | 258 | 918 | 28% | 4% | 10.09% |
| **(731.0, 768.0]** | 686 | 220 | 906 | 24% | 4% | 9.08% |
| **(768.0, 797.0]** | 715 | 218 | 933 | 23% | 4% | 8.19% |
| **(797.0, 820.0]** | 728 | 159 | 887 | 18% | 4% | 7.22% |
| **(820.0, 841.0]** | 801 | 134 | 935 | 14% | 4% | 6.52% |
| **(841.0, 859.0]** | 777 | 124 | 901 | 14% | 4% | 5.95% |
| **(859.0, 873.0]** | 793 | 94 | 887 | 11% | 4% | 5.35% |
| **(873.0, 886.0]** | 800 | 109 | 909 | 12% | 4% | 4.92% |
| **(886.0, 898.0]** | 876 | 88 | 964 | 9% | 4% | 4.28% |
| **(898.0, 908.0]** | 851 | 73 | 924 | 8% | 4% | 3.77% |
| **(908.0, 917.0]** | 793 | 56 | 849 | 7% | 4% | 3.30% |
| **(917.0, 926.0]** | 855 | 54 | 909 | 6% | 4% | 2.91% |
| **(926.0, 935.0]** | 929 | 40 | 969 | 4% | 4% | 2.48% |
| **(935.0, 942.0]** | 813 | 37 | 850 | 4% | 4% | 2.19% |
| **(942.0, 950.0]** | 931 | 33 | 964 | 3% | 4% | 1.78% |
| **(950.0, 957.0]** | 893 | 18 | 911 | 2% | 4% | 1.34% |
| **(957.0, 965.0]** | 864 | 17 | 881 | 2% | 4% | 1.12% |
| **(965.0, 975.0]** | 900 | 10 | 910 | 1% | 4% | 0.72% |
| **(975.0, 996.0]** | 881 | 3 | 884 | 0% | 4% | 0.34% |

**Segmento Reactiva**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DECILES** | **BUENOS** | **MALOS** | **TOTAL** | **TM** | **%** | **TM ACUMULADA** |
| **(4.999, 97.0]** | 25 | 914 | 939 | 97% | 4% | 30.86% |
| **(97.0, 183.0]** | 92 | 834 | 926 | 90% | 4% | 28.07% |
| **(183.0, 274.0]** | 184 | 745 | 929 | 80% | 4% | 25.38% |
| **(274.0, 372.0]** | 280 | 648 | 928 | 70% | 4% | 22.89% |
| **(372.0, 464.0]** | 362 | 575 | 937 | 61% | 4% | 20.66% |
| **(464.0, 538.0]** | 456 | 475 | 931 | 51% | 4% | 18.61% |
| **(538.0, 597.0]** | 522 | 410 | 932 | 44% | 4% | 16.90% |
| **(597.0, 645.28]** | 569 | 351 | 920 | 38% | 4% | 15.39% |
| **(645.28, 685.0]** | 603 | 341 | 944 | 36% | 4% | 14.07% |
| **(685.0, 716.0]** | 632 | 315 | 947 | 33% | 4% | 12.67% |
| **(716.0, 743.0]** | 646 | 267 | 913 | 29% | 4% | 11.27% |
| **(743.0, 766.0]** | 731 | 194 | 925 | 21% | 4% | 10.01% |
| **(766.0, 787.0]** | 756 | 184 | 940 | 20% | 4% | 9.17% |
| **(787.0, 808.0]** | 781 | 153 | 934 | 16% | 4% | 8.29% |
| **(808.0, 828.0]** | 780 | 154 | 934 | 16% | 4% | 7.55% |
| **(828.0, 846.0]** | 777 | 129 | 906 | 14% | 4% | 6.65% |
| **(846.0, 865.0]** | 828 | 102 | 930 | 11% | 4% | 5.83% |
| **(865.0, 884.0]** | 875 | 107 | 982 | 11% | 4% | 5.19% |
| **(884.0, 899.0]** | 820 | 97 | 917 | 11% | 4% | 4.32% |
| **(899.0, 914.0]** | 869 | 62 | 931 | 7% | 4% | 3.28% |
| **(914.0, 929.0]** | 851 | 45 | 896 | 5% | 4% | 2.60% |
| **(929.0, 945.0]** | 924 | 29 | 953 | 3% | 4% | 2.02% |
| **(945.0, 960.0]** | 878 | 25 | 903 | 3% | 4% | 1.67% |
| **(960.0, 975.0]** | 938 | 19 | 957 | 2% | 4% | 1.13% |
| **(975.0, 998.0]** | 899 | 2 | 901 | 0% | 4% | 0.22% |

**Segmento Reprogramado**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DECILES** | **BUENOS** | **MALOS** | **TOTAL** | **TM** | **%** | **TM ACUMULADA** |
| **(5.999, 329.0]** | 195 | 1028 | 1223 | 84% | 4% | 16.91% |
| **(329.0, 507.0]** | 448 | 775 | 1223 | 63% | 4% | 14.11% |
| **(507.0, 627.88]** | 666 | 554 | 1220 | 45% | 4% | 11.97% |
| **(627.88, 701.0]** | 801 | 435 | 1236 | 35% | 4% | 10.45% |
| **(701.0, 751.0]** | 864 | 360 | 1224 | 29% | 4% | 9.26% |
| **(751.0, 785.0]** | 914 | 299 | 1213 | 25% | 4% | 8.25% |
| **(785.0, 810.0]** | 963 | 286 | 1249 | 23% | 4% | 7.39% |
| **(810.0, 832.0]** | 1003 | 212 | 1215 | 17% | 4% | 6.51% |
| **(832.0, 850.0]** | 1055 | 190 | 1245 | 15% | 4% | 5.87% |
| **(850.0, 865.0]** | 1023 | 183 | 1206 | 15% | 4% | 5.27% |
| **(865.0, 879.0]** | 1076 | 145 | 1221 | 12% | 4% | 4.62% |
| **(879.0, 892.0]** | 1124 | 120 | 1244 | 10% | 4% | 4.10% |
| **(892.0, 903.0]** | 1168 | 115 | 1283 | 9% | 4% | 3.66% |
| **(903.0, 913.0]** | 1131 | 80 | 1211 | 7% | 4% | 3.20% |
| **(913.0, 922.0]** | 1096 | 80 | 1176 | 7% | 4% | 2.89% |
| **(922.0, 931.0]** | 1204 | 77 | 1281 | 6% | 4% | 2.51% |
| **(931.0, 939.0]** | 1153 | 59 | 1212 | 5% | 4% | 2.10% |
| **(939.0, 947.0]** | 1231 | 47 | 1278 | 4% | 4% | 1.75% |
| **(947.0, 954.0]** | 1170 | 40 | 1210 | 3% | 4% | 1.45% |
| **(954.0, 960.0]** | 1105 | 17 | 1122 | 2% | 4% | 1.14% |
| **(960.0, 967.0]** | 1323 | 18 | 1341 | 1% | 4% | 1.07% |
| **(967.0, 973.0]** | 1172 | 26 | 1198 | 2% | 4% | 1.00% |
| **(973.0, 980.0]** | 1202 | 13 | 1215 | 1% | 4% | 0.60% |
| **(980.0, 987.0]** | 1175 | 7 | 1182 | 1% | 4% | 0.35% |
| **(987.0, 999.0]** | 1121 | 1 | 1122 | 0% | 4% | 0.09% |

### 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

## ANEXO

|  |  |
| --- | --- |
|  | Modelo |
| Modelo vigente |  |
| Periodos |  |
| Diccionario de variables |  |
| Códigos |  |
| Objeto del Modelo XGBOOST |  |

## RESPONSABILIDAD DE ÁREA Y ÓRGANOS INTERVENIENTES

En cada una de las etapas descritas, se han tenido reuniones periódicas entre las siguientes áreas:

* Áreas de Riesgos:
  1. Modelamiento de Riesgos.
  2. Departamento de Riesgo BPE.

En dichas reuniones las decisiones de ejecución del desarrollo han sido acordadas por todas las partes.

Así mismo, debido a las limitaciones de data y la coyuntura post-pandemia se ha acordado revisar periódicamente el modelo y evaluar un reentrenamiento cuando se tenga disponible un mayor volumen de periodos con tasas de morosidad estables.

1. 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. [↑](#footnote-ref-1)
2. Se está aplicando el forzado en las variables *min\_umdmneg6umv2, max\_a\_pcsldresjud\_3umv2, x6um\_pc\_sldrct\_nonorv2 y num\_adcta\_9um* para poder tener un comportamiento monótono más claro a la hora de la interpretabilidad de la variable. [↑](#footnote-ref-2)
3. Se está aplicando el forzado en la variable *ant\_negociov2* para poder tener un comportamiento monótono más claro a la hora de la interpretabilidad de la variable. [↑](#footnote-ref-3)
4. Se está aplicando el forzado en las variables *group\_rt\_3avgtrpsvin6* y *rt\_6pstotal12umv2* para poder tener un comportamiento monótono más claro a la hora de la interpretabilidad de la variable. [↑](#footnote-ref-4)