Prueba Analítica: Riesgo de Crédito

DICAGI 2021

# Introducción

Uno de los modelos clásicos en la gestión de crédito es el modelo de predicción de incumplimiento del pago de obligaciones. Este modelo, tiene como objetivo calcular la probabilidad de que un cliente -en los próximos n meses, llegue a una altura de mora determinada-.

Con la probabilidad que entrega el modelo, usualmente se clasifican los clientes en grupos de riesgo. Con esos grupos de riesgo se generan políticas diferenciadas. Por ejemplo: algunos productos se ofrecen sólo a algunos grupos, mientras que otros productos pueden tener un apetito de riesgo más alto (aceptar grupos más riesgosos) o más bajo (aceptar menos grupos). Otros productos, por ejemplo, pueden tener políticas de aprobación de clientes de riesgo alto, sin embargo, piden requisitos adicionales como por ejemplo codeudores y/o garantías.

Esta prueba analítica tiene como objetivo el desarrollo de un modelo de riesgo, para un segmento particular de clientes. Este segmento de clientes -en general- no hace un uso intensivo de los canales del banco. Por otro lado, son clientes con experiencia crediticia, es decir: alguna entidad financiera del país ya les ha desembolsado algún crédito.

El modelo desarrollado debe clasificar a la población en los ocho grupos que se muestran en la Tabla 1. En esta tabla se incluye el límite inferior y superior de probabilidad de incumplimiento esperada por grupo. Esto quiere decir que para los clientes clasificados en el grupo T2, se espera que tengan una tasa de incumplimiento mínima de 1%+ y máxima del 1.5%.

|  |  |
| --- | --- |
| Grupo de Riesgo | Límites de Probabilidad |
| T1 | [0%, 1%] |
| T2 | (1%, 1.5%] |
| T3 | (1.5%, 3%] |
| T4 | (3%, 4.5%] |
| T5 | (4.5%, 8%] |
| T6 | (8%, 15%] |
| T7 | (15%, 30%] |
| T8 | (30%, 100%] |

Tabla – Grupos de riesgo con su respectivo límite de probabilidad

Es importante que cada grupo no se sobrepase ni por encima (entregando una tasa de incumplimiento superior) ni por debajo; ya que esto implicaría que se está corriendo un riesgo mayor del esperado, o que se está dejando clientes buenos por fuera de mercado objetivo, respectivamente.

De modo que un buen modelo es aquel que:

* Asigna tasas de incumplimiento “en rango” para cada grupo de riesgo. Es decir, se busca maximizar la proporción de clientes en rango.
* Distribuye la población en los distintos grupos de riesgo. Esto significa que un modelo que sólo clasifica la población en dos grupos es peor que un modelo que los clasifica en tres.

# Objetivo y Restricciones

Se debe desarrollar una solución analítica que clasifique los clientes en ocho categorías de riesgo. La tasa de incumplimiento de cada grupo debe estar dentro de los rangos definidos en la Tabla 1.

El modelo debe ser optimizado para los clientes marcados con la columna tipo\_cliente = ‘objetivo’. Los clientes marcados con tipo\_cliente = ‘adicion’ pueden ser utilizados para entrenar, sin embargo, las métricas de las soluciones se revisarán sólo para los clientes objetivo.

La base tiene información desde enero de 2018 hasta febrero de 2019. El modelo sólo debe entrenarse con información de 2018. La información de 201901 debe usarse para probar. La información de 201902 será la que utilizaremos nosotros para verificar los resultados de su modelo.

# Datos

Se entregarán bases de datos en formato csv (separador pipe) con n columnas. En los adjuntos se encuentra el diccionario de variables en donde se hace una descripción de cada una de las columnas de la tabla.

Sin embargo, a continuación, se dan detalles adicionales sobre algunos campos que merecen una explicación adicional:

|  |  |
| --- | --- |
| Nombre del Campo | Descripción |
| tipo\_cliente | Esta es una variable categórica con dos categorías:  objetivo: Estos son los clientes de interés en donde se desea utilizar el modelo.  adicion: Estos son clientes que comparten algunas características con los clientes objetivo, sin embargo, tienen orígenes diferentes lo que significa que son no son poblaciones iguales. |
| default | Esta es la variable respuesta:  Toma el valor de uno cuando el cliente incumplió el pago de su obligación.  Toma el valor de cero cuando el cliente cumplió con el pago. |

Los tres archivos entregados contienen la siguiente información:

|  |  |
| --- | --- |
| Nombre del Archivo | Descripción |
| base\_train | Esta es la información que debe utilizarse para entrenar el modelo. Corresponde a desembolsos realizados en 2018.  Incluye las variables explicativas y la variable respuesta. |
| base\_validacion | Esta base la podrá utilizar para probar los resultados de su modelo. Esta información no debe incluirse en el entrenamiento del modelo.  Corresponde a desembolsos realizados en 201901.  Incluye las variables explicativas y la variable respuesta. |
| base\_prueba | Esta base será con la que evaluaremos su modelo. Corresponde a desembolsos de 201902.  Incluye sólo las variables explicativas. |

# Tiempos y Entregables

El propósito de la prueba es medir sus capacidades de análisis y desarrollo de modelos analíticos. La idea es que no le dedique más de 15 horas en total, incluyendo el tiempo para documentar lo que hizo.

Es posible usar cualquier herramienta que quiera (Python, R, Spark, SAS Guide/Miner, SPSS, etc.), y cualquier recurso del internet, pero no se permite consultar directamente con otras personas por ningún medio.

Puede realizar los supuestos que considere necesarios. No es necesario utilizar todos los datos o todas las variables. Esto depende de la forma en que usted aborde el problema. No hay una solución única. Inclusive puede darse el caso en que no se tenga un modelo viable.

Los entregables son los siguientes:

1. Documento en donde explique el proceso que siguió para resolver la prueba. Incluyendo detalle sobre todas las etapas que haya seguido para solucionar esta prueba. Si alguna fase del desarrollo analítico no se incluye en este documento, al evaluar la prueba se asumirá que no se hizo. De manera opcional nos podría hacer saber qué otros datos o atributos adicionaría idealmente al conjunto de datos, para un modelo analítico más efectivo. Aquí, tenga en cuenta la factibilidad y el costo de obtener esos datos.
2. Bases de prueba calificadas con la predicción del grupo de riesgo.

Esta base debe ser entregada en formato csv. El grupo de riesgo debe dejarse en una columna con nombre: grupo\_riesgo.

Todos los clientes entregados deben ser calificados con una de las siguientes categorías [t1, t2, t3, t4, t5, t6, t7, t8]. Ningún cliente debe quedar con predicción nula, vacía o con una categoría diferente a las ocho mencionadas.

1. Códigos documentados que respalden el ejercicio analítico.

Estos son los archivos de código con comentarios en caso de usar un lenguaje de programación convencional o el archivo de proyecto que incluya documentación, en caso de usar SAS Miner, Azure ML Studio, u otra herramienta parecida. Puede enviarlos en una carpeta comprimida o darnos acceso a un repositorio de github.

1. Con el uso del modelo en mente, diseñe un sistema de manera teórica que bosqueje una solución que permita hacer disponibles los resultados de su modelo analítico y que sean fácilmente consumibles por servicios externos, páginas web, servicio móvil, etc. No se tiene que desarrollar. Desarrollar una aplicación o sistema de información no da ningún punto extra y no será tenida en cuenta para la calificación total por lo que recomendamos no desarrollarla sino únicamente elaborar el bosquejo y describir la estrategia.

# Evaluación

La métrica para evaluar este modelo será el porcentaje de población en rango, este porcentaje entre más alto mejor será su desempeño en la prueba. Para esto, se utilizará la información de desembolsos de 201902. De estos desembolsos sólo nosotros conocemos la variable respuesta.