Prueba de conocimiento analítico

DiCAGI 2018

El propósito de la prueba es medir sus capacidades de análisis y desarrollo de modelos predictivos. 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 (R, Python, SAS Guide/Miner, SPSS, etc.), y cualquier recurso del internet, pero no se permite consultar directamente con otras personas por ningún medio.

1. Introducción

El área de conciliación con el cliente es un área con gran impacto sobre el cliente y la organización, ya que es la encargada de velar para que los indicadores de cartera vencida y el valor de provisiones se encuentren dentro de márgenes sanos para la organización. Parte del trabajo de esta área consiste en no permitir que los clientes vencidos en sus cuotas o pagos se incrementen a través del tiempo, para lo cual, es importante la prevención del vencimiento de las obligaciones en las edades de mora más tempranas. Es por esto que el área de conciliación propuso anteriormente que una vez la obligación entra en mora, ésta sea gestionada enviando mensajes de texto (SMS), correos electrónicos y llamadas personalizadas. Esto ha incrementado considerablemente el costo de gestión, por lo que el área de conciliación quiere desarrollar un modelo analítico de auto-cura.

El modelo de auto-cura tiene como objetivo predecir de manera oportuna, qué cliente que recién entra en mora, pagará su obligación, en menos de 15 días sin realizarle ningún tipo de gestión o solo realizando gestiones indirectas. Entiéndase gestión indirecta como gestiones de tipo masivo, es decir, SMS o correo electrónico (las cuales tienen un valor muy bajo). Este modelo permitirá

determinar a qué clientes gestionar desde etapas más tempranas y, por consiguiente, disminuir los costos en gestiones innecesarias a clientes que realmente pagarían sin gestiones directas. En otras palabras, auto-cura se define como aquellos clientes que de manera oportuna (y sin contacto directo) se ponen al día en menos de 15 días (a partir del momento en que entran en mora).

2. Definición de la población objetivo y variable respuesta

El área de conciliación ha decidido que se comience con el desarrollo del modelo analítico por sectores, comenzando con uno de los segmentos más críticos en el tema de morosidad, el segmento de pyme pequeña. El modelo de auto-cura posee una variable respuesta u objetivo, que se define como:

- Toma el valor de 1 si la obligación es pagada en menos de 15 días (inclusive) sin realizarle gestiones o solamente gestiones de formato indirecto.
- Toma el valor de 0 en cualquier otro caso.

3. Instrucciones importantes

- El archivo adjunto *Base_entrenamiento.csv* contiene información sobre muchas obligaciones en diferentes ciclos de pago, esto significa que una obligación puede aparecer varias veces en la base, pero una única vez por un ciclo de pago, la columna *llave* es un identificador único de la obligación del cliente por cada ciclo de facturación, *anhomes_ciclo* es una columna que indica el año y el mes del ciclo de facturación que está asociado a la obligación del cliente que se encuentra dentro de la variable *llave*.
- En la *Base_entrenamiento.csv* existe una columna llamada *y_auto_cura* dicha variable, es una variable dicótoma, que indica el valor de si la obligación del cliente para ese ciclo de facturación se auto-curó o no, de acuerdo a la definición de variable respuesta de la sección anterior.

- El archivo *Descrip_Variables.xlsx* contiene la descripción o definición de cada una de las columnas de la base de datos *Base_entrenamiento.csv.*

- El objetivo es que usted desarrolle un modelo analítico que permita, a partir de los datos del archivo *Base_entrenamiento.csv* predecir si una obligación asociada a un cliente se auto-curará o no en cualquier ciclo de facturación.

- El archivo adjunto *Base_prueba.csv* contiene exactamente las mismas columnas del archivo *Base_entrenamiento.csv,* exceptuando la columna *y_auto_cura.*

4. Entregables

- Se debe entregar un archivo *Base_prueba_evaluado.csv* con las columnas de *llave* y *probabilidad*, la columna probabilidad es un valor real (es decir fraccionario) entre 0 y 1, dónde 0 indicará que no hay ninguna posibilidad de que la obligación se auto cure, 1 indicará que la obligación se auto curará con total certeza y 0.5, por ejemplo, indicaría que existe igual probabilidad de que se auto cure o no, es decir, la columna probabilidad debe entregar la probabilidad de que la variable respuesta definida en la sección 2 sea 1.

- La columna probabilidad **debe** contener un valor real entre 0 y 1 (inclusive) **para todos y cada uno** de los registros. No aceptaremos valores nulos, NaNs, N/A, N/D, vacíos, o mensajes de texto como por ejemplo: “datos incompletos”. Por favor haga todo lo posible por conservar el formato del archivo (csv separado por comas, no otro carácter; el orden de las columnas; la línea de encabezado etc. El orden de las filas no es crítico).

- También nos debe entregar la implementación de su modelo (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).

- Un archivo de texto (en .txt , .doc, .html, .rmd, .md o .pdf) que contenga una descripción más o menos detallada del proceso que siguió para generar el modelo (incluyendo exploración, transformaciones de variables, selección de variables, etc.) y luego generar sus predicciones de auto-cura.

- De manera opcional nos podría hacer saber qué otros datos o atributos de las obligaciones o clientes, de su comportamiento frente al banco añadiría idealmente al conjunto de datos, para un modelo predictivo más efectivo. Aquí, tenga en cuenta la factibilidad y el costo de obtener esos datos.

- Ahora, es posible que llegue a la conclusión de que no es posible desarrollar un buen modelo predictivo a partir de la información proporcionada o dada la calidad de la misma. Si este es el caso, queremos evaluar el mejor modelo que pueda producir y también que nos dé una sustentación de esa conclusión.

5. Evaluación

La métrica para evaluar el modelo será el **AUC** entre el valor real de auto-cura para cada obligación en un ciclo de facturación, que sólo nosotros conocemos, y el valor predicho por su modelo y consignado en el archivo *Base_prueba_evaluado.csv* por medio de la columna *probabilidad*.

Adicionalmente se evaluará integralmente el informe entregado: el análisis de las variables, su transformación, proceso de selección, y cualquier componente analítico que haya sido útil para la construcción del modelo.