**Equipo 30**

**A01793999 - Nathalia Prada**

**A01793933 - Christian Suárez Gil**

1. ¿Qué datos considero más importantes? ¿Por qué?

Para entender cuales datos son más importantes hay que tener claridad sobre es el objetivo del estudio, en el caso de los datos presentados para el ejercicio el objetivo es predecir desde la perspectiva de gestión de riesgos la probabilidad de incumplimiento de pago de los clientes en Taiwán.

Con este objetivo en mente pienso que los datos más importantes presentados son:

* La variable de respuesta “Y”: pago por defecto (Sí = 1, No = 0).

En cuanto a las 23 variables explicativas para esta respuesta “Y”, pienso que es difícil determinar cuáles son las más importantes sin antes realizar un estudio estadístico inicial, pero intuitivamente considero que los datos más relevantes para predecir la probabilidad de incumplimiento de un pago son los siguientes:

* X1: Monto del crédito otorgado (dólar NT): incluye tanto el crédito de consumo individual como su crédito familiar (complementario).
* X4: Estado civil (1 = casado; 2 = soltero; 3 = otros).
* X5: Edad (año).
* X6 - X11: Historial de pagos pasados. Seguimiento de los registros de pagos mensuales pasados ​​(de abril a septiembre de 2005)
  + La escala de medición del estado de amortización es:
    - -1 = pagar debidamente;
    - 1 = retraso en el pago de un mes;
    - 2 = retraso en el pago de dos meses
    - 8 = retraso en el pago de ocho meses
    - 9 = retraso en el pago de nueve meses o más.
* X12-X17: Importe del estado de cuenta (dólar NT).

1. ¿Se eliminaron o reemplazaron datos nulos? ¿Qué se hizo y por qué?

Se decidió eliminar los valores nulos debido a que se encontró que 42 filas de las 30000 que componen el data set contienen al menos un valor vacío, estas 42 filas representan tan solo el 0.14% de la muestra por lo que se decide eliminarlas en lugar de reemplazarlas debido a que el reemplazo de estos valores puede generar problema como los siguientes:

* No preservar la relación entre las variables.
* Subestimar el error estándar

1. ¿Es necesario ordenar los datos para el análisis? Sí / No / ¿Por qué?

No considero que sea necesario ordenar los datos para el análisis debido a que cada fila tiene un registro por cliente y esta es la relación de datos que se debe mantener durante el análisis independientemente del orden dado a dichos clientes.

1. ¿Existen problemas de formato que deban solucionar antes del proceso de modelado? Sí / No / Por qué.

Se realizaron reemplazos en la data “X3: Educación” para cambiar los datos no registrados en el rango descrito en para la data: (1 = posgrado; 2 = universidad; 3 = secundaria; 4 = otros), el valor 0 se reemplazó por 1 = posgrado y los valores 5 y 6 se reemplazaron por 4 = otros.

También en las columnas X6 - X11: Historial de pagos pasados, se encontraron valores de -2 que no hacen parte de la descripción del rango,

1. ¿Qué ajustes se realizaron en el proceso de limpieza de datos (agregar, integrar, eliminar, modificar registros (filas), cambiar atributos (columnas)?

En el proceso de limpieza de datos se realizaron los siguientes pasos:

1. Lectura del archivo en un dataframe.
2. Cambiar el nombre de las columnas pro un nombre más fácil de leer, de la siguiente manera:
   * X1: credit
   * X2: gender
   * X3: education
   * X4: marital
   * X5: age
   * X6 - X11: Historial de pagos pasados (History of Past Payment hpp)
     + X6: hpp Sep 2005
     + X7: hpp Aug 2005
     + X8: hpp Jul 2005
     + X9: hpp Jun 2005
     + X10: hpp May 2005
     + X11: hpp Apr 2005
   * X12-X17: Importe del estado de cuenta (dólar NT) (Amount of bill statement (NT dollar) abs)
     + X12: abs Sep 2005
     + X13: abs Aug 2005
     + X14: abs Jul 2005
     + X15: abs Jun 2005
     + X16: abs May 2005
     + X17: abs Apr 2005
   * X18-X23: Monto del pago anterior (dólar NT). (Amount of previous payment (NT dollar) app)
     + X18: app Sep 2005
     + X19: app Aug 2005
     + X20: app Jul 2005
     + X21: app Jun 2005
     + X22: app May 2005
     + X23: app Apr 2005
3. Identificar la cantidad de valores faltantes en cada columna, la columna con más valores faltantes es “X15: abs Jun 2005” con 17 valores faltantes, la única columna sin valores faltantes es “X1: credit”.
4. Utilizar la librería missingno para visualizar los valores faltantes en el data set: los valores faltantes no representan visualmente ninguna diferencia en el data set.
5. Describir la información: utilizando la función describe:
   * Se encuentra existen registros de nivel educativo de 1 a 6 niveles, pero en la descripción de la data solo se mencionan 4 niveles. Se realizaron reemplazos en la data “X3: Educación” para cambiar los datos no registrados en el rango descrito para la data: (1 = posgrado; 2 = universidad; 3 = secundaria; 4 = otros), el valor 0 se reemplazó por 1 = posgrado y los valores 5 y 6 se reemplazaron por 4 = otros.
   * Se encuentra que en las casillas X6 - X11: Historial de pagos pasados (hpp) el mínimo valor de la escala es -2 pero en la descripción de la data el mínimo valor debería ser -1, por lo tanto, se decide reemplazar el valor -2 y 0 con -1.
6. Usar el parámetro thresh en la función dropna para identificar el número de filas que tienen un determinado número de columnas vacías. Experimentando con esa función se construyó la siguiente tabla:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| thresh | Filas que cumplen con el thresh | Filas descartadas (que no cumplen con el thresh) | Porcentaje de filas descartadas | Número de columnas vacías |
| 24 | 29958 | 42 | 0.14% | 0 |
| 23 | 29976 | 24 | 0.08% | 1 |
| 22 | 29988 | 12 | 0.04% | 2 |
| 21 | 29988 | 12 | 0.04% | 3 |
| 18 | 29989 | 11 | 0.04% | 6 |
| 14 | 29991 | 9 | 0.03% | 10 |
| 13 | 29992 | 8 | 0.03% | 11 |
| 11 | 29997 | 3 | 0.01% | 13 |
| 10 | 29998 | 2 | 0.01% | 14 |
| 7 | 29999 | 1 | 0.00% | 17 |
| 3 | 30000 | 0 | 0.00% | 21 |
| 0 | 30000 | 0 | 0.00% | 24 |

Con esta información se puede ver que 42 filas tienen al menos una columna vacía, y 3 filas tienen 21 columnas vacías.

1. Determinar el porcentaje de filas de todo el dataframe que tienen al menos un valor vacío: 42 filas con valores vacíos corresponden al 0.1399% de la cantidad de filas del dataframe.
2. Debido a que las filas con valores vacíos representan menos del 1% de la muestra se decide eliminar todas las filas con valores vacíos.
3. Se realizaron reemplazos en la data “X3: Educación” para cambiar los datos no registrados en el rango descrito para la data: (1 = posgrado; 2 = universidad; 3 = secundaria; 4 = otros), el valor 0 se reemplazó por 1 = posgrado y los valores 5 y 6 se reemplazaron por 4 = otros.
4. se decide reemplazar el valor -2 con -1, para las columnas X6 - X11: Historial de pagos pasados.