CREDIT ONE

REPORTE CLIENTES QUE NO PAGARÁN SUS CREDITO

# Problema:

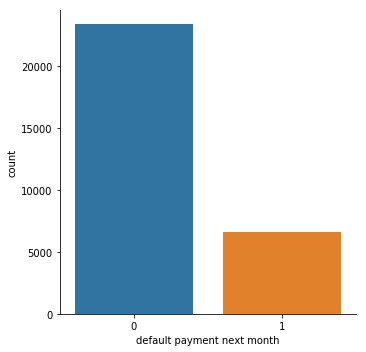
Incremento en la tasa de no pago de clientes lo cual es malo para CREDIT ONE, dado que su negocio es pre aprobar clientes en primer lugar. Esto pone en riesgo el negocio de CREDIT ONE.

# Preguntas

1. ¿Cómo se puede asegura que un cliente puede o está dispuesto a pagar su crédito?
2. ¿Podemos aprobar clientes con alta certeza?

# Base de datos

Se dispuso de 30 mil observaciones del comportamiento de pago de los clientes, estando clasificados como clientes que Pagan= 0, y que no Pagan = 1:



De las 30 mil observaciones, 77% corresponden a clientes que pagan (etiquetados como 0) y el resto 22% a clientes que no pagan (etiquetados como 1). Aquí se ve como los datos no están balanceados porque existen más observaciones de clientes que pagan que de clientes que no pagan, casi una relación de 4:1.

Variables Explicativas:

Se cuenta con un total de 23 variables explicativas:

1. Demográficas 3:

|  |
| --- |
| **Variable** |
| Estado Civil (“MARRIAGE”) |
| Sexo (“SEX”) |
| Nivel Educativo (“EDUCATION”) |

1. Sobre el uso de la tarjeta de crédito se dispone de 20 variables, que en realidad presentan la información para 3 aspectos, para cada uno de los meses comprendidos entre abril y setiembre de 2005.

|  |
| --- |
| **Variable** |
| Historia Pasada de Pago (“PAY\_X”) |
| Estado de Cuenta mensual  (“BILL\_AMTX”) |
| Pago Mensual  (“PAY\_AMTX”) |

# Tratamiento del Problema

La cuestión es como mejorar los parámetros que procuran determinar si un cliente atenderá sus obligaciones financieras a futuro, es decir, etiquetar a un cliente como potencial pagador o potencial no pagador.

Sin embargo, lo más importante, es establecer un nivel de probabilidad, como un límite a partir del cual un cliente sea clasificado como Pagador y por lo tanto, cada nuevo cliente que supere ese umbral sea etiquetado como Pagador.

Dado que se conoce la clasificación actual de los clientes de la base de datos, se está frente a un problema de clasificación que se cataloga como “Aprendizaje Supervisado”.

Se revisan los datos y aprovechando que ya están codificados en términos numéricos, se procede a revisar su composición y a modificar aquellos que por su tipo deben ser definidos como categóricos (EDUCACIÓN, SEXO, ESTADO CIVIL, COMPORTAMIENTO DE PAGO HISTÓRICO). De igual forma se estudia y analizan las variables numéricas como EDAD, Monto del Estado de Cuenta, Pago Mensual. En este caso se procedió a convertir la variable EDAD en una variable categórica para lo cual primero se realizó una división de la misma en subconjuntos por edad.

Terminada esa etapa se procedió a probar 4 posibles clasificadores:

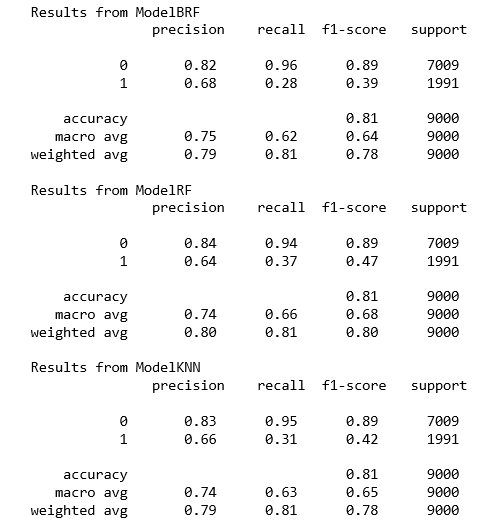
1. Balanced RAndom Forest
2. Random Forest
3. KNeighbors Classifier
4. Gradient Boosting Classifier

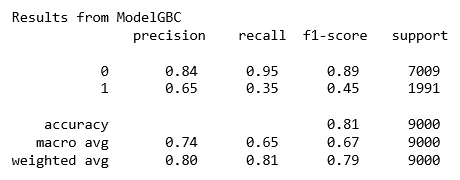
# Comparación de Clasificadores y escogencia del Modelo

Cada uno de los modelos fue optimizado utilizando una técnica que permite automatizar la cantidad de parámetros asociados a cada modelo en procura de obtener la mayor Precisión (Accuracy)

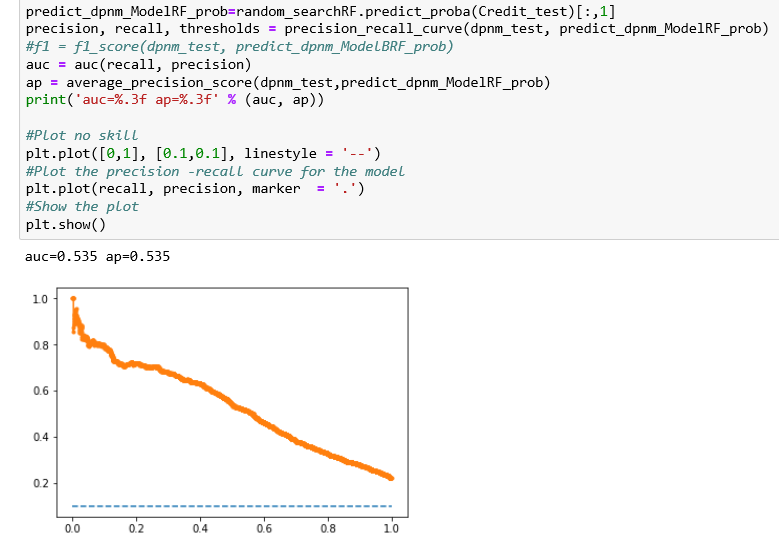
Nota: el primero de la lista se incluyó al final procurando mejorar los indicadores que permiten clasificar los modelos por su desempeño, dado que al estar los datos en una relación cercana a 4 observaciones de clientes que Pagan por cliente que No Pagan, existe un desbalance que debe ser corregido.

Los resultados obtenidos fueron los siguientes:





Por el problema de desbalance de los datos se prefirió utilizar la métrica conocida como F1-score siendo la más alta para el modelo Random Forest, aunque se debe decir que no son satisfactorios sus datos para tratar de reducir el número de potenciales clientes que podrían ser catalogados como Pagadores, cuando en realidad revisten el riesgo de ser lo contrario. Lo cual se puede ver en la siguiente curva:



Esta gráfica permite ver que muy rápidamente el modelo clasifica falsos positivos, antes de alcanzar una Precisión de 0.1 ya ha clasificado mal a clientes que son No pagadores. Lo que no es lo mejor para CREDIT ONE. El modelo BalancedRandomForest que se supone está hecho para compensar bases de datos desbalanceadas presentó niveles de R1 menores al modelo RandomForestClassifier.

# Obtención de la probabilidad

No obstante lo mencionado anteriormente, para el modelo Random Forest se calcularon las probabilidades de ser clasificado entre cada una de las posibles etiquetas de este ejercicio binario (0 o 1).

Dado el sesgo que tiene el modelo, con niveles de probabilidad mayores o iguales a 25% de 1100 clientes que se sabe que son No Pagadores, se estarían clasificando como No Pagadores 740 y como Pagadores 360.

Pero con un nivel de probabilidad mayor a igual a 35.5% de 844 clientes que se sabe que son No Pagadores, el modelo estarían clasificando bien 709 y como Pagadores 135.

Para un nivel de probabilidad mayor a igual a 45.5% de 647 clientes que se sabe que son No Pagadores, el modelo estarían clasificando bien 608 y como Pagadores 39.



# Variables explicativas más importantes

