**Tema 5**

**Task 2: Prepare and explore the data**

**Francisco Paniagua Barrantes**

**Preguntas**

How do you ensure that customers can/will pay their loans?

Can we approve customers with high certainty?

No se puede asegurar con completa certeza que un cliente no caerá en Default, pero se pude seleccionar con un nivel de certeza los clientes que tienen una alta probabilidad de no caer en Default.

El modelo desarrollado inicialmente nos permite seleccionar con un 42% de exactitud de que el resultado sea correcto para seleccionar los clientes que no caerán en Default.

**Hallazgos**

Se identifico que al aumentar el monto del pago disminuye el Default.

Se observa que los créditos entre 0 y 200000 dólares son los que presentan mayor tendencia al default, por lo tanto, se recomienda revisar los filtros para los créditos de montos de entre 0 y 200000 dólares.

Las variables que presentan la mayor importancia y asociación con el default se presentan en la figura 1.

Las variables que mayor importancia para pronosticar si un cliente caerá o no en Default corresponden a PAY\_0, AGE, LIMIT\_BAL, PAY\_AMT1, PAY\_AMT2, PAY\_AMT3, PAY\_AMT5, PAY\_AMT6,PAY\_AMT4. Las demás variables se eliminaron y no se utilizaron para generar el pronóstico.

Al analizar los indicadores, se selecciono como indicador clave Recall, esto porque motivo de que es preferible falsos positivos que falsos negativos.

Los resultados de los indicadores obtenidos para los diferentes clasificadores probados es el que se muestra en el cuadro 1.

El modelo seleccionado corresponde a un decisión tree con criterion = 'gini', max\_depth = 200, esto por motivo de que con este algoritmo se obtuvo el mayor recall, como se muestra en el cuadro 1.

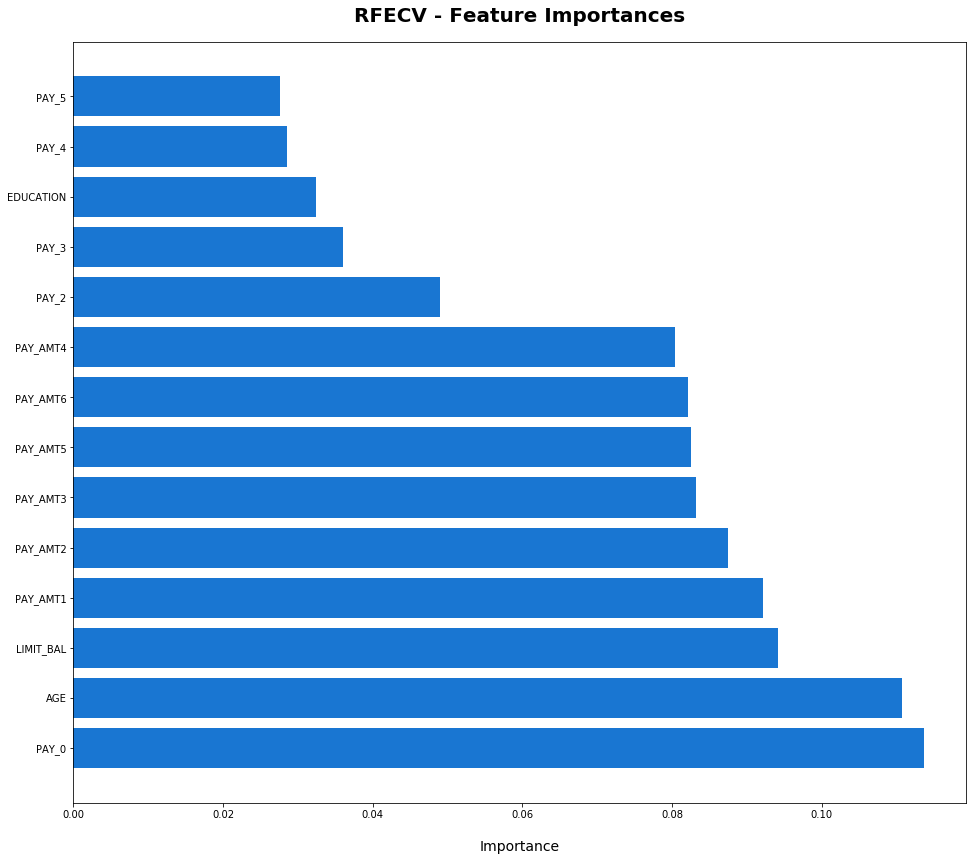


Figura 1. Importancia de las variables



Cuadro 1. Indicadores de los modelos desarrollados

**Recomendaciones**

Se recomienda desarrollar más la efectividad del algoritmo para realizar el pronóstico, pero se requiere para esto un equipo con una mayor capacidad de procesamiento y dos semanas más para mejorarlo.

Se observo que que se obtiene mejores resultados al establecer los parámetros de forma empírica o manual en vez de utilizar el grid.