Reporte

CreditOne

—

Modelo Predictivo

—

Michael Fallas León

RESUMEN

Durante el último año, la empresa Credit One ha presenciado un incremento en el número de clientes que han incumplido los préstamos que han obtenido de varios de sus socios, y Credit One, al ofrecer el servicio de calificación crediticia, podría arriesgarse a perder negocios si el problema no se resuelve.Por lo tanto, han solicitado, al equipo de ciencia de datos, que presenten un diseño e implementen una solución creativa y empíricamente sólida. Su implementación se da utilizando la herramienta de Python y algunas de sus bibliotecas. El desarrollo del proyecto se lleva a cabo por medio de algunos datos históricos de sus clientes, el cual consta de un total de 30 000 observaciones.

Además, se muestran aspectos destacados del análisis y la visualización de los datos facilitados por la empresa de Credit One. Por lo tanto, en éste documento se destaca la información que podría ser relevante para la empresa, con el fin de buscar soluciones a su problema actual y entender el comportamiento de los clientes.

Finalmente, se muestran los modelos predictivos construidos, así como, el rendimiento de los mismo, con el fin de seleccionar el modelo que nos determine con mejores parámetros de rendimiento, si el crédito se debe o no otorgar al cliente y, con ello minimizar el riesgo de otorgárselo a clientes que tengan una probabilidad alta de incumplir con el pago.

Con la realización del proyecto, se desea dar respuesta a las siguientes preguntas:

- ¿Cómo se puede asegurar que los clientes puedan/vayan a pagar los créditos?
- ¿Se puede aprobar clientes con alta certeza?

En la presente sección se muestra un resumen del análisis obtenido de la exploración de los datos (EDA), donde se detallan aspectos que destacan en cada uno de los atributos, así como la visualización de estos. Con el fin de entender influencias de los atributos, al momento de definir si el cliente pagará o no la cuota del crédito.

A continuación, se muestran visualizaciones para determinados atributos, así como el análisis de estos:

Edad del cliente
Haciendo un enfoque en el atributo de edad, no se logra evidenciar que éste tenga una fuerte influencia al momento de determinar si el cliente realizará o no el pago el próximo mes. Por lo tanto, se concluye que la edad no influye en el pago de las facturas de cada crédito (ver figura 1).

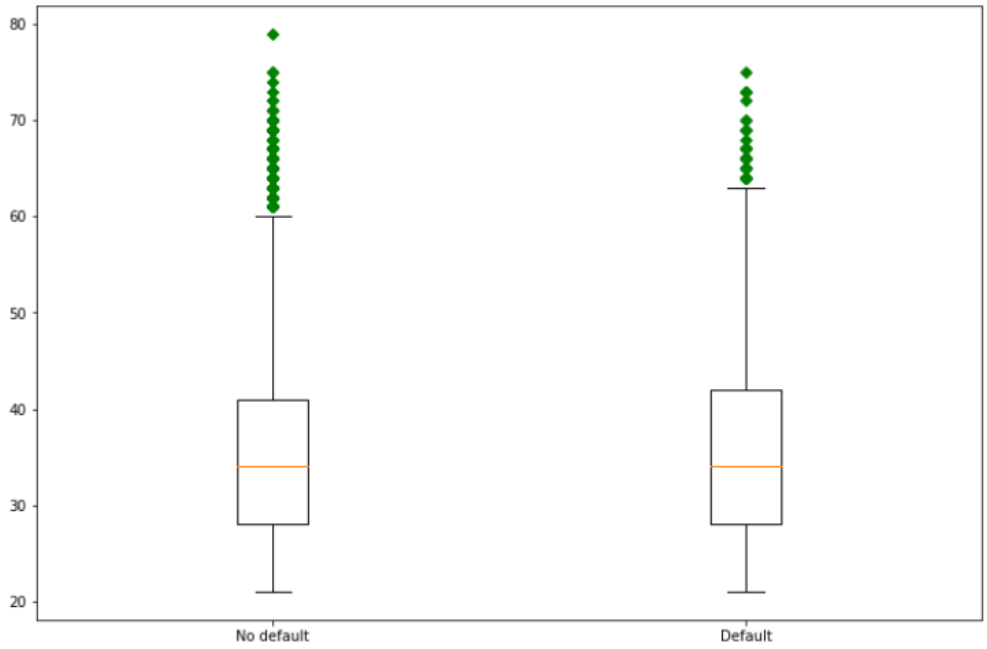

Figura 1. Gráfico de cajas: comportamiento de pago vs la edad del cliente

Sin embargo, se logra evidenciar que la gran mayoría de los clientes poseen una edad que va desde los 30 años hasta los 40 años aproximadamente.

Límite de crédito del cliente
Lo más destacable del análisis realizado con éste atributo es que el porcentaje de clientes que no realizarán el pago del crédito el próximo mes tiende a disminuir conforme aumenta el límite del crédito del cliente; en éste caso, el porcentaje de clientes que incumplen cuando el límite de crédito es menor a los $100 000 es de 29.5% y,

disminuye hasta un 13% cuando los clientes poseen un límite de crédito mayor a $500 000.

Otro aspecto interesante es que, se puede observar que no se dispone de clientes cuyo límite de crédito se encuentre dentro de los $800 000 y $900 000; por lo tanto, es importante considerar incrementar la cantidad de observaciones con el fin de tener datos para todas las opciones posibles.

Lo anterior se puede evidenciar en el siguiente gráfico de barras, donde se muestran los límites del balance en rangos de $100 000 cada uno, a excepción del último nivel que va desde $900 000 a los $100 000 000, y el porcentaje que representan los clientes que se espera que paguen o no en el siguiente mes.

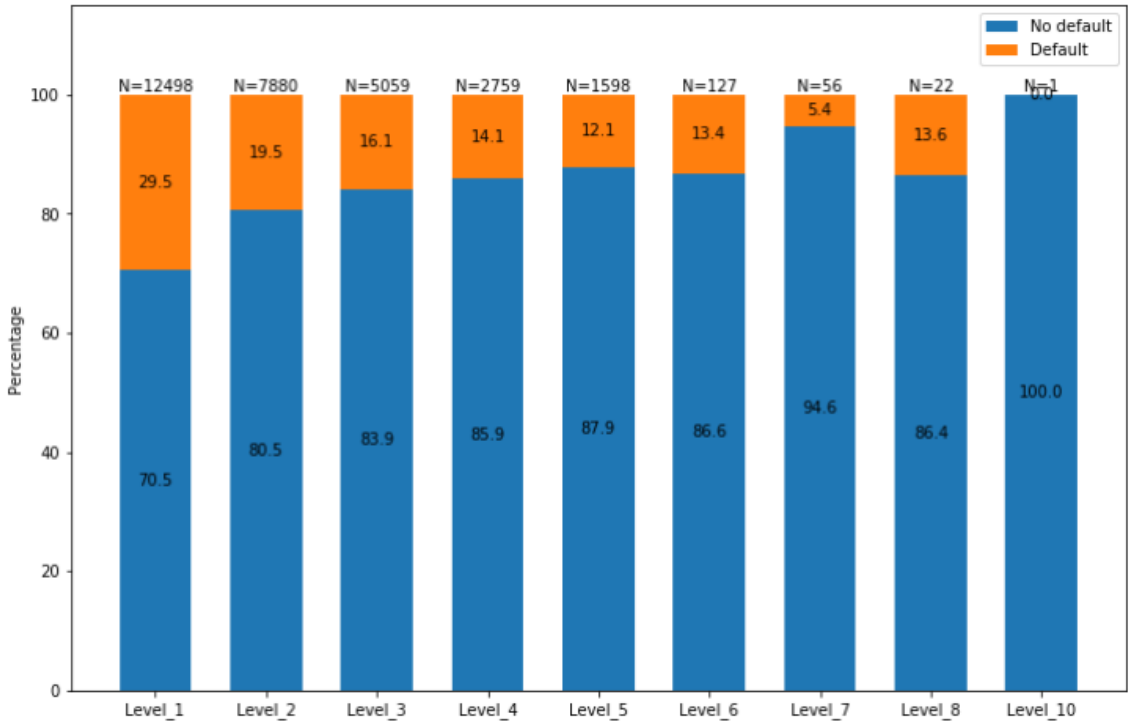

Figura 2. Gráfico de barras: Límite de balances vs el porcentaje de clientes que pagarán y que no pagarán

Nivel de educación del cliente

Al principio se pensaba que el porcentaje de clientes que incumplirán con el pago del crédito iba a disminuir conforme sea más alto su nivel de escolaridad; sin embargo, se observa un comportamiento totalmente opuesto, el porcentaje de los clientes que poseen educación universitaria es de 23.7%, ligeramente superado por los clientes que terminaron el colegio, mientras que en la categoría de "otros" (en los que consideramos títulos técnicos, capacitaciones, ningún estudio, etc.), el porcentaje de clientes que se estima no pagará el próximo mes es de 7.1% (ver figura 3).

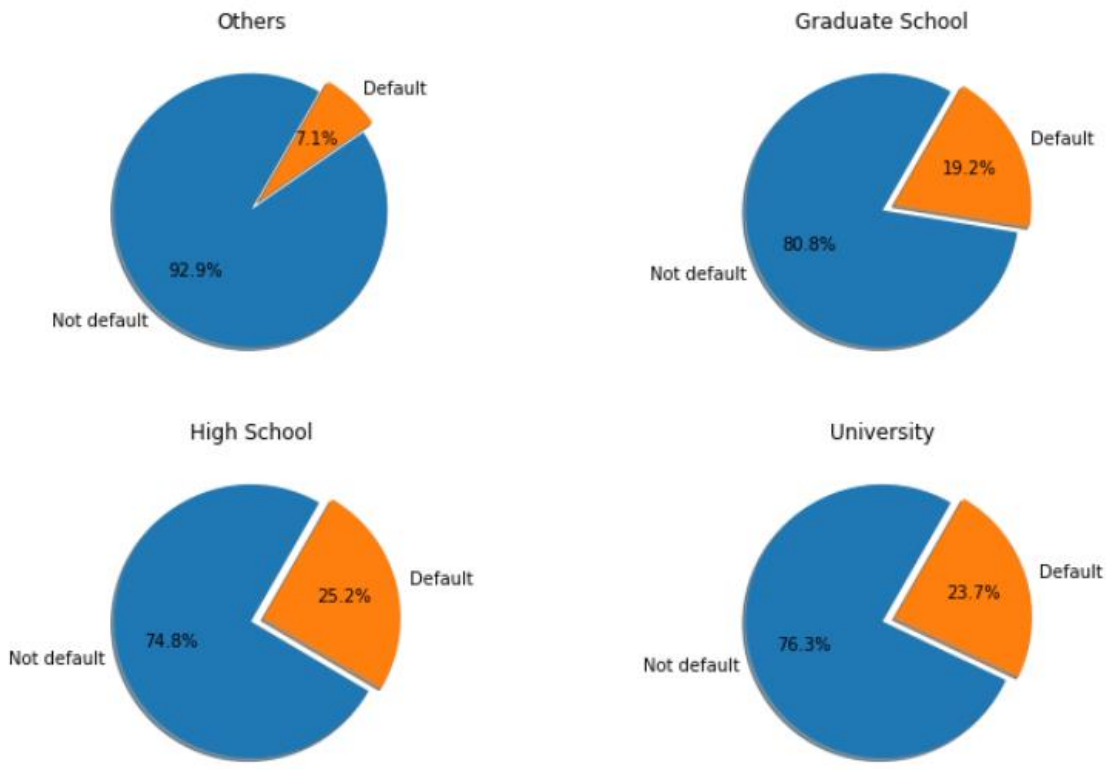

Figura 3. Gráfico de pie: porcentaje de clientes que pagarán o no el próximo mes separado por nivel de educación del cliente

Es importante resaltar que la cantidad de observaciones en ésta última categoría es mucho menor a las demás, por lo que, con el fin de aumentar la confiabilidad de los resultados, se necesita aumentar las observaciones en ésta categoría. Lo cual se puede evidenciar en la siguiente tabla.

Tabla 1. Cantidad de clientes que puede que paguen o no según el nivel de educación

Education	Non default	Default
University	10 700	3 330
Graduate school	8 549	2 0366
High School	3 680	1 237
Other	435	33

Estado civil del cliente

Para éste atributo, cabe resaltar que menos de 400 clientes se encuentran en las categorías de “divorciado(a)” y “otros”, por lo que el mayor porcentaje de las observaciones corresponden a clientes que se encuentran casados(as) o solteros(as). Tomando en cuenta solamente estas dos últimas categorías, no se observa una clara diferencia entre los clientes que se espera que paguen o no el crédito, ya que los porcentajes que muestran el incumplimiento del pago del crédito es de 23.5% para aquellos clientes cuyo estado civil es casado(a), y un 20.9% para los clientes con estado civil de soltero(a). Para los otros dos estados civiles, divorciado(a) y otros, los porcentajes de clientes que posiblemente no paguen el próximo mes es de 26%

(aumento) y 9.3% (disminuye). Estos resultados se pueden observar en el siguiente gráfico.

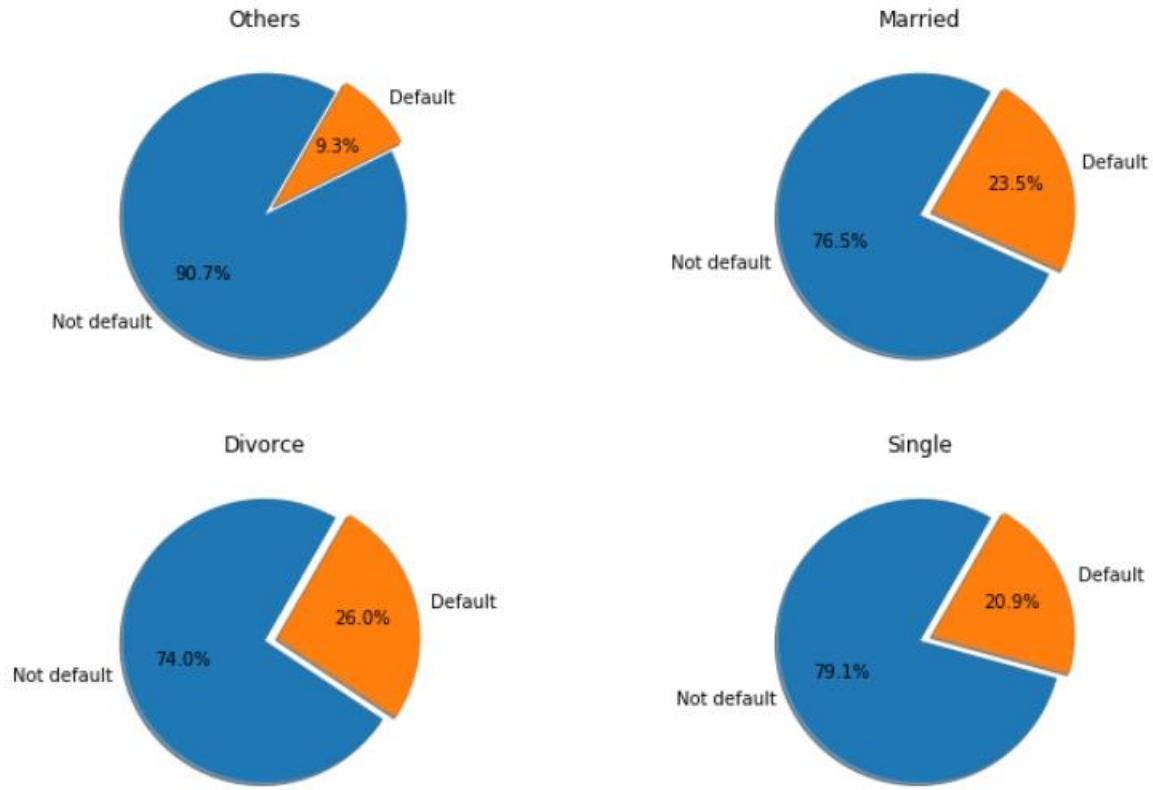

Figura 4. Gráfico de pie: porcentaje de clientes que pagarán o no el próximo mes separado por el estado civil del cliente

Sin embargo, la cantidad de clientes en las últimas categorías son muy pocos como se menciona anteriormente, por lo que se recomienda buscar más observaciones para ambas categorías (ver tabla 2).

Tabla 2. Cantidad de clientes que puede que paguen o no según según su estado civil

Marriage	Non default	Default
Married	10 453	3 206
Single	12 623	3 341
Divorce	239	84
Other	49	5

Historial de pago en un semestre del cliente

Se realzó el análisis del historial crediticio de los clientes, tomando en cuenta la cantidad de facturas en las cuales el cliente no efectuó ningún pago, lo cual se determinó por la resta entre la cantidad de meses en los que el pago fue de $0 y la cantidad de meses en los que el monto de la factura era de $0. En e siguiente gráfico se muestra los resultados obtenidos.

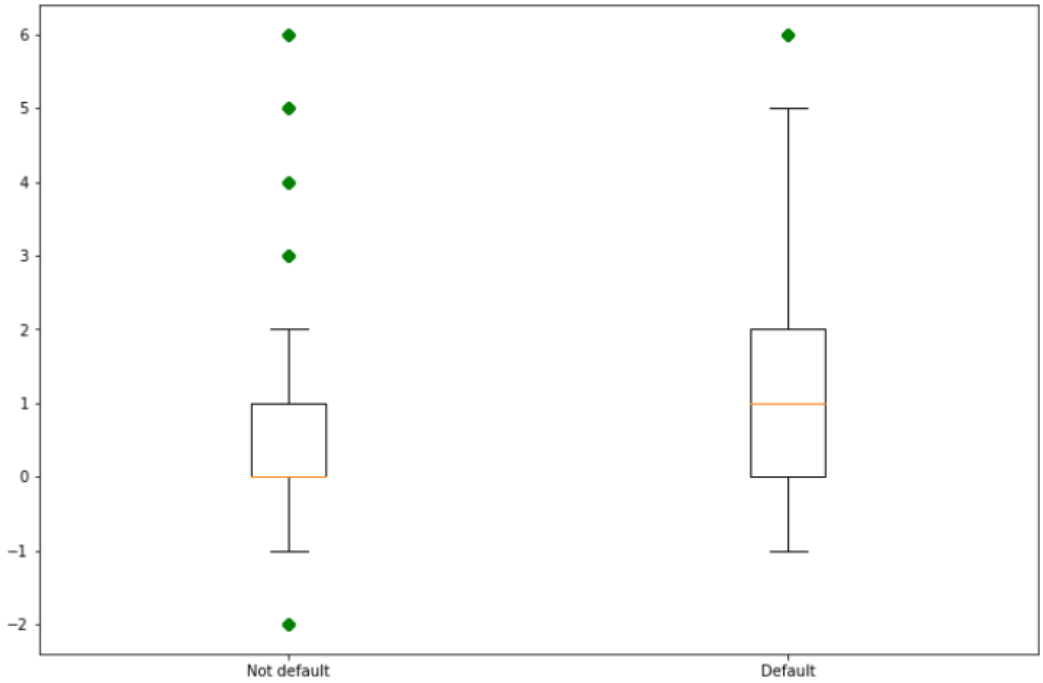

Figura 5. Gráfico de cajas: comportamiento de pago vs la cantidad de pagos no realizados

De lo anterior se puede determinar que el 50% de los clientes que posiblemente pagarán el mes siguiente no dejaron de pagar ninguna de sus facturas en los seis meses que se disponen, el 75% dejaron de pagar solamente una factura, y para el total, como máximo, estos clientes dejaron de pagar dos de sus facturas (25% restante). Además, se observan muy pocas observaciones sobrepasan estos 2 meses sin pago.

Por otra parte, para el caso en el que se predice que no se realizará el pago por parte de los clientes en el mes siguiente, el 50% de estos no pago en al menos una ocasión, el 75% de los clientes no pagaron en al menos dos ocasiones, mientras que para el 25% restante, los valores van desde 2 a 5 meses sin pagar.

Por lo tanto, se puede determinar que si el cliente tiene varios meses sin pagar (al menos 2 de los últimos 6) alguna factura del crédito superior a $0, muy posiblemente éste no pague la próxima factura (si ésta es superior a 0$).

Modelos Construidos

Se construyeron modelos basados en los datos facilitados por la empresa, utilizando tres tipos de algoritmos con el fin de determinar la solución que prediga con un mejor rendimiento si el cliente efectuará o no el pago de la cuota el siguiente mes.

Cabe destacar que para la selección de atributos se utiliza PCA (Principal Component Analysis) con el fin de reducir la dimensionalidad de los datos, lo cual es aplicado solamente a los atributos numéricos que en este caso corresponden al historial crediticio de los últimos 6 meses. Además, se eliminan los atributos que no presentarán algún beneficio o no ofrece aporte alguno en nuestro proyecto, como es el caso del atributo *ID*, y, con el fin de evitar problemas en el futuro por el tema de discriminación, se elimina el atributo *SEX*.

También, se toma la decisión de discretizar algunos datos continuos; es decir, colocar los valores de los atributos en depósitos para que exista un número limitado de estados posibles, los cuales son tratados como si fueran valores ordenados y discretos. Para los datos a disposición, se decide discretizar el atributo *AGE* en 4 depósitos, que abarca lo siguiente:

- Young adults (1): 15 a 29 años.
- Adults (2): 30 a 44 años.
- Middle-aged adults (3): 45 a 59 años.
- Older adults (4): 60 a 99 años.

El otro atributo modificado es *LIMIT_BAL* en 10 depósitos, con límites cada $100 000, a excepción del último depósito que abarca préstamos muy elevados ya que son muy pocos.

Los algoritmos utilizados para la construcción de los modelos son los siguientes: *Random Forest*, *Support Vector Machine* y *K-Nearest Neighbors*. La siguiente tabla muestra los parámetros de rendimiento en las predicciones para los modelos obtenidos.

Tabla 3. Parámetros de rendimiento de los modelos

Model	Accuracy	Kappa
modelRF	81.6%	0.36
modelSVM	78.4%	0.0
ModelKNN	78.5%	0.10

Observando los resultados anteriores, el modelo con mejor rendimiento corresponde al modelo que utliza el algoritmo de *Random Forest Classifier*, ya que posee un 81.6% de exactitud y el valor de kappa más alto.

A continuación, se muestra la importancia que le da el modelo seleccionado a cada uno de los atributos:

1. feature 13: PC3 (0.123410)
2. feature 10: PC0 (0.123131)
3. feature 11: PC1 (0.117059)
4. feature 12: PC2 (0.116422)
5. feature 14: PC4 (0.116359)
6. feature 4: PAY_0 (0.106774)
7. feature 5: PAY_2 (0.052598)
8. feature 6: PAY_3 (0.037276)
9. feature 1: EDUCATION (0.033255)
10. feature 0: LIMIT_BAL (0.032570)
11. feature 3: AGE (0.032558)
12. feature 7: PAY_4 (0.030197)
13. feature 9: PAY_6 (0.030029)
14. feature 8: PAY_5 (0.025503)
15. feature 2: MARRIAGE (0.022858)

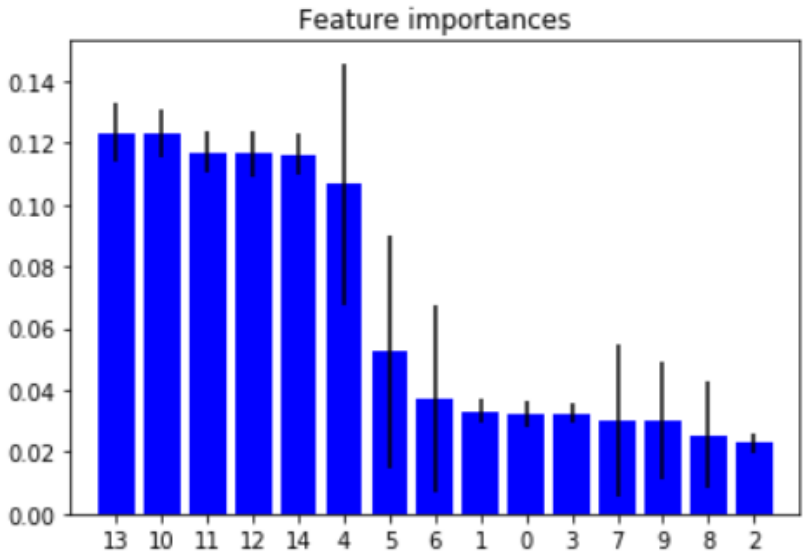

Figura 6. Importancia de las características

Conclusiones

Por medio del modelo desarrollado es posible predecir con una precisión de aproximadamente un 81% si el cliente realizará o no el pago del crédito. Por lo tanto, respondiendo a la pregunta planteada por Credit One, no es posible asegurarse que los clientes puedan o vayan a pagar los créditos. Sin embargo, es posible predecir con un porcentaje de confiabilidad alto si el cliente efectuará o no el pago, esto con el fin de no seleccionar a clientes que no lleguen a pagar el crédito.

Del mismo modo, con los atributos disponibles, la aprobación de los créditos a clientes se puede aprobar con el mismo porcentaje de certeza. Sin embargo, si se desea mejorar el rendimiento del modelo, se recomienda disponer de más datos e incluso considerar otros atributos como el salario de los clientes; esto con el fin de que el modelo busque patrones que le faciliten la predicción y así pueda aumentar la exactitud de este.

Es importante mencionar que los atributos más importantes para el modelo fueron las variables pertenecientes al historial crediticio de los últimos seis meses de cada uno de los clientes. Por lo tanto, antes de la aprobación del crédito, el enfoque principal será el historial crediticio, por encima de la edad, estado civil o educación del cliente, ya que la importancia de estas últimas para definir el pago o no de las cuotas del préstamo es muy baja.