Reporte Final: Riesgo Crediticio

Jairo Ordóñez Guardado
Credit One

Contexto

Hace poco más de un año hemos notado un incremento en el número de personas que no pagan sus obligaciones de crédito adquiridas con nuestros clientes. La función de Credit One en el proceso, es brindar a nuestros socios comerciales la mejor recomendación posible respecto a cuáles clientes son elegibles para otorgarles un crédito.

Si nuestras recomendaciones no mejoran, el impacto financiero que están sufriendo nuestros clientes puede acrecentarse y convertirse en una situación crítica que les puede llevar al cierre de operaciones. Esto directamente nos afecta porque la confianza y credibilidad en nuestro trabajo se está viendo seriamente cuestionada.

En este punto la pregunta clave planteada es **¿La persona que llega a solicitar un préstamo pagará SÍ o NO el crédito?**

En este reporte abordaremos los puntos clave en torno a este análisis.

Objetivos

Basándonos en el contexto anterior y la pregunta clave del negocio hay dos objetivos principales:

- Analizar qué factores influyen para que un cliente pague o no su crédito.
- Desarrollar una herramienta que nos permita tener una buena predicción de los clientes que sí pagarían un préstamo vs los que no lo pagarían.

Hallazgos y Resultados

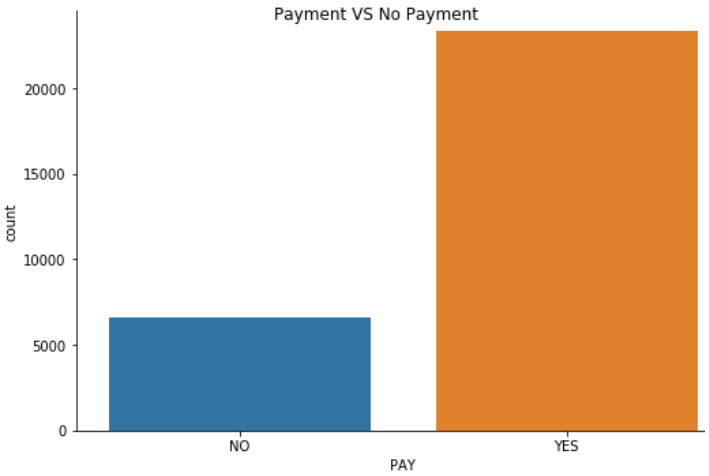
Antes de abordar los resultados de este estudio consideramos importante dejar claro que, para términos de entendimiento, decidimos reacomodar la variable respuesta de los datos. Por lo tanto, para efectos de las visualizaciones y detalles que se mencionen a continuación, cuando nos referimos a la categoría YES, estaremos haciendo énfasis en los clientes que SÍ pagan su préstamo, en contraste con los de la categoría NO (no pagan).

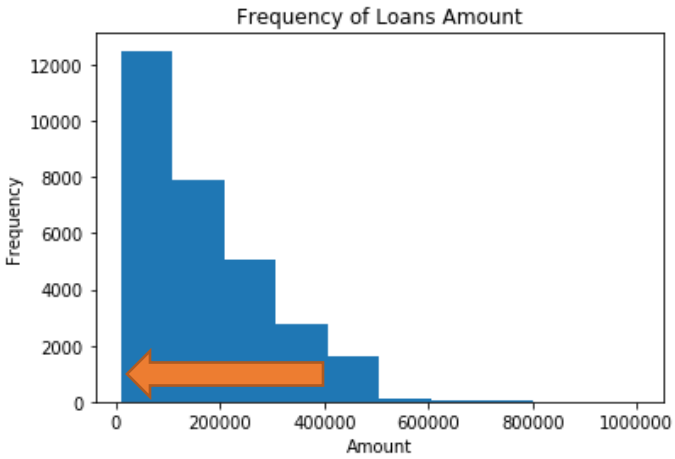
Definitivamente es un caso un poco complejo principalmente por dos factores:

- Desde el planteamiento del caso mencionamos que hacía falta un poco más de información para desarrollar mejores relaciones y un modelo, sobre todo la carencia de un dato tan clave en temas de aprobaciones de crédito como lo es el ingreso económico. Aunque este dato no te asegure una respuesta afirmativa por sí mismo, sí te dice si un cliente potencial tiene capacidad de pago para el monto del préstamo solicitado.
- Indirectamente el caso tiene relación con el comportamiento humano y esto es algo que se sale de control para cualquier análisis de datos. No sabemos y no es posible controlar los hábitos de gastos de las personas lo cual afecta directamente el compromiso de un cliente con sus obligaciones crediticias.

Al observar los datos hay un aspecto que nos interesa saber de primera mano: clientes que sí pagan vs clientes que no pagan. Aquí nos encontramos con una proporción cercana al 80/20, SÍ y NO respectivamente.

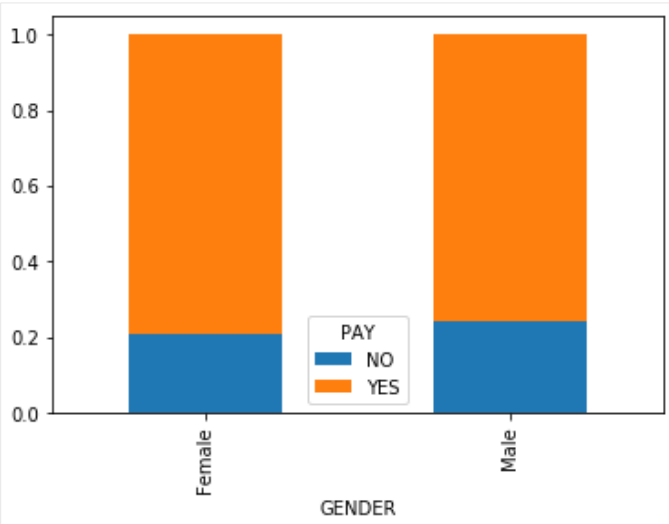
La gráfica siguiente nos lo muestra y a su vez podemos pensar que tenemos un conjunto de datos desbalanceado lo cual es un factor a considerar posteriormente en el modelado, habría una tendencia a entrenar y predecir mejor los clientes que sí pagan vs los que no pagan, aunque esto no sería del todo negativo para el caso en cuestión.

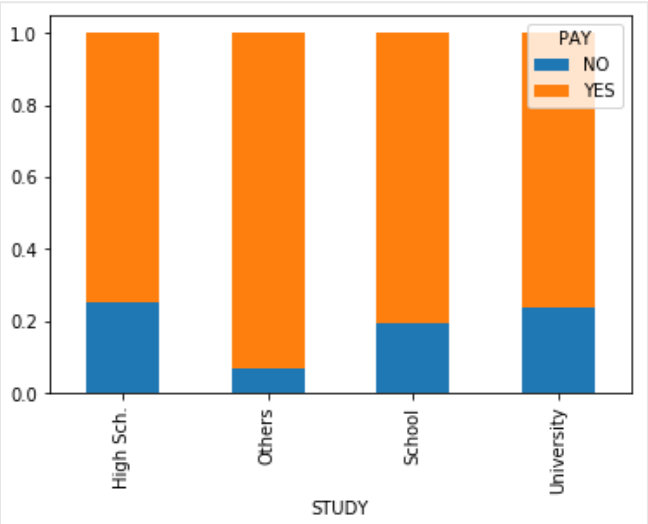




Otra información que es importante conocer es cómo se distribuyen los préstamos en torno a sus montos, aquí fue posible identificar que el 94% de los créditos otorgados no superan el monto de los $400mil y que este límite representa casi el 83% de todo el dinero adeudado en el conjunto de datos para el análisis.

Si exploramos factores individuales que inciden en el pago o no de una deuda tenemos varios escenarios. Por ejemplo, a nivel de SEXO la primera gráfica nos muestra que el 79% de las mujeres SÍ paga, mientras que para los hombres el porcentaje de pago es de un 76%. Esto nos diría que las clientes femeninas son de alguna manera más confiables.

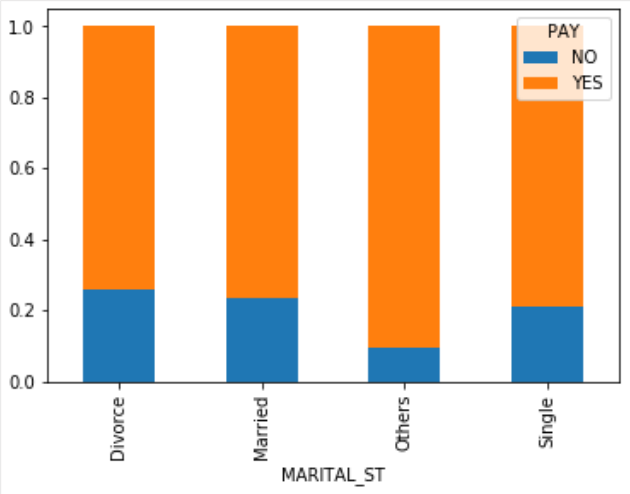


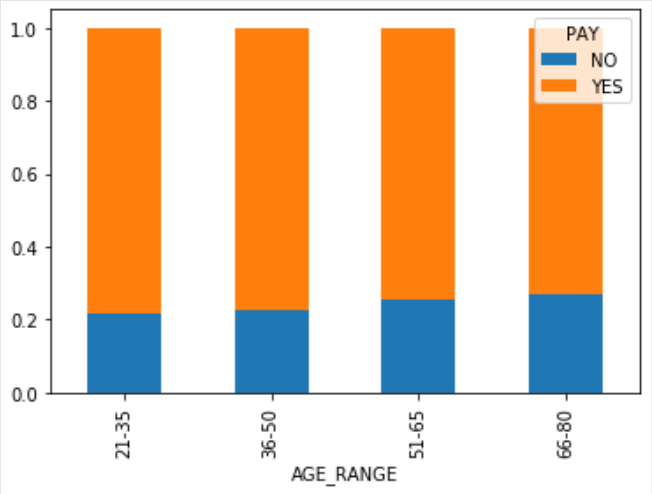


Por otro lado, en términos de nivel educativo OTROS estudios son los que mejor proporción de pago registran con un 93% SÍ. No obstante, esta categoría es donde menos clientes tenemos. El grueso de clientes se concentra en las categorías Universitarios y Escuela, lo interesante es que los clientes con estudio escolar presentan una proporción positiva de pago mejor a los universitarios, un 81% vs un 76%. Este dato es llamativo porque podríamos suponer que los clientes con universidad tienen mejores ingresos y por ende una mejor capacidad de pago, pero esto no se está reflejando en los datos.

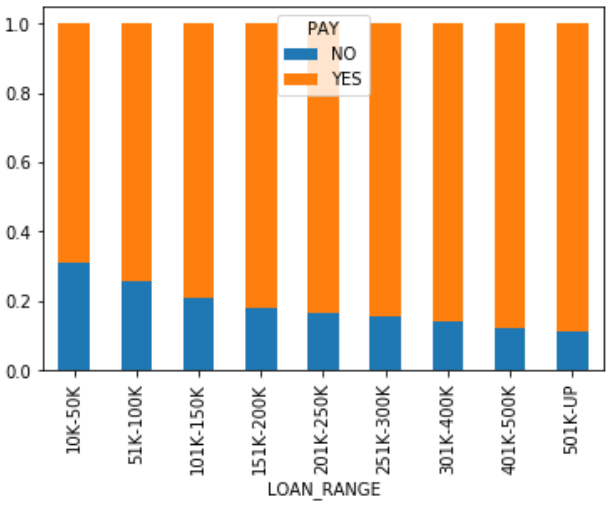
Si observamos el estado civil de los clientes, hay dos grupos principales en población, Solteros y Casados y de estos destacan más los solteros que presentan un porcentaje de pago del 79% vs un 77% los casados. Es decir, de alguna forma resultaría mejor otorgar préstamos a clientes solteros.

En términos de edad decidimos agrupar en 4 segmentos y podemos decir que la mayoría de clientes no superan los 50 años (92% de los clientes en los datos) y hasta este límite de edad la proporción de SÍ pago es de un 79% menores de 35 años y 77% entre 36 y 50 años. A mayor edad existe una leve tendencia a aumentar el porcentaje de NO pago.



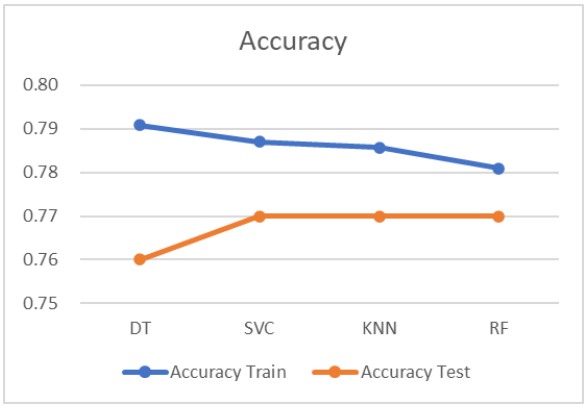


Por último, es interesante observar como los préstamos de menor monto son los que más alta proporción de NO PAGO presentan, pensábamos que iba a ser todo lo contrario. A medida que el monto del préstamo es mayor, la proporción positiva de pago va aumentando. Consideramos que esto puede obedecer a que préstamos de sumas altas de dinero, tienen mayores consecuencias al no ser canceladas. A pesar de esto, no debemos olvidar el punto mencionado al inicio, la mayoría de préstamos no superan el monto de los $400K.



Modelos

Debido a que el problema central es responder si un cliente potencial pagará o no un préstamo, consideramos que el modelo principal debería centrarse únicamente en aquellas características que se conocen antes de otorgar un préstamo. Sin bien es cierto, el conjunto de datos con el que se trabajó el caso contiene variables sobre datos mensuales de pagos y estados de cuenta históricos de préstamos ya efectuados, estas variables no tienen valor cuando se valorará a un cliente por primera vez.



Por esta razón el primer modelo lo limitamos únicamente a datos que podríamos denominar “conocidos”. Es decir, aquellos datos que obtenemos cuando el cliente llega al área de servicio a solicitar un crédito por primera vez, estos datos son el monto del crédito solicitado y variables demográficas del cliente: sexo, estado civil, edad y educación. Con este set de variables trazamos cuatro modelos diferentes obteniendo los resultados de exactitud mostrados en el gráfico al lado.

Como se puede observar, los valores de exactitud anduvieron muy parecidos entre los modelos y también entre la fase de entrenamiento y prueba lo cual nos habla de consistencia en los modelos. La exactitud osciló entre el 76% y 79% entre ambas etapas. A pesar de estas similitudes en exactitud nos inclinamos por KNN. Básicamente descartamos los demás por factores como costo computacional (SVC) con un tiempo de entrenamiento mucho mayor, RF solo predijo valores YES al observar su matriz de confusión y el Decision Tree además de predecir menos valores NO que el KNN, también presentó la mayor diferencia entre las métricas del entrenamiento y testing.

Las métricas generales y matriz de confusión del clasificador KNN que seleccionamos son las siguientes.

	precision	recall	f1-score	support
NO	0.38	0.06	0.11	2029
YES	0.78	0.97	0.87	6971
accuracy			0.77	9000
macro avg	0.58	0.52	0.49	9000
weighted avg	0.69	0.77	0.69	9000

Predicted	NO	YES	All
True			
NO	127	1902	2029
YES	204	6767	6971
All	331	8669	9000

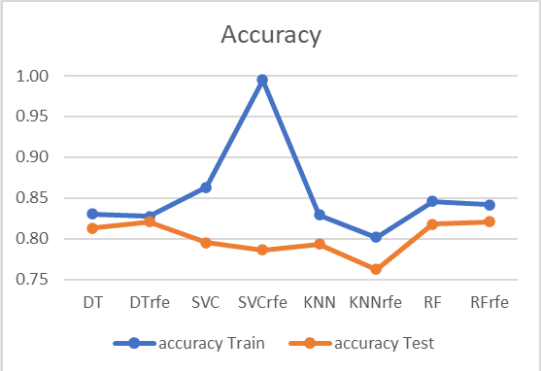
Como se puede ver, la exactitud general en el testing fue del 77%, no es muy alta como se quisiera, pero ronda lo que los cuatro modelos promediaron bajo el planteamiento que indicamos al inicio. Este dato nos diría que el modelo es capaz de predecir correctamente casi en un 80% si un cliente pagaría o no un préstamo solicitado. No obstante, y como se ve a nivel de precisión y recall, estas predicciones son más favorables cuando se trata de SÍ pago, esto puede ser producto de un data set desbalanceado como mencionamos anteriormente, aunque tener un alto valor para YES sigue siendo un punto positivo ya que es el objetivo principal del negocio.

Otros Modelos

Llevamos a cabo otras pruebas con más variables considerando aquellos datos que corresponden a créditos que ya han sido efectuados y que tienen valores mensuales. Básicamente probamos dos conjuntos de datos diferentes con cuatro clasificadores.

El primero de estos, al que llamaremos Experimento 1 incluye los siguientes datos: el monto del préstamo, las variables demográficas y las variables categóricas del estado de los pagos de los últimos meses ('LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_STA_SET', 'PAY_STA_AGO', 'PAY_STA_JUL', 'PAY_STA_JUN', 'PAY_STA_MAY', 'PAY_STA_APR').

El otro grupo de modelos (Experimento 2) fueron realizados con un conjunto de datos sugeridos mediante RFE (Recursive Feature Elimination) y nos definió las siguientes características: 'LIMIT_BAL', 'AGE', 'PAY_STA_SET', 'BILL_AMT_SET', 'BILL_AMT_AGO', 'BILL_AMT_JUL', 'BILL_AMT_JUN', 'BILL_AMT_MAY', 'BILL_AMT_APR', 'PAY_AMT_SET'.



En esta ocasión seleccionamos el clasificador RF (Random Forest) del experimento 1 pues consideramos que tiene buenas métricas de exactitud y consistentes entre lo evaluado en entrenamiento (85%) como en prueba (82%). además presentó valores significativos en precisión y recall tanto para respuestas SÍ como NO.

Cabe mencionar que el Arbol de Decisión con RFE también tuvo métricas aceptables también.

A continuación las métricas generales del Random Forest:

	precision	recall	f1-score	support
NO	0.69	0.37	0.48	2045
YES	0.84	0.95	0.89	6955
accuracy			0.82	9000
macro avg	0.76	0.66	0.68	9000
weighted avg	0.80	0.82	0.80	9000

Predicted	NO	YES	All
True			
NO	750	1295	2045
YES	344	6611	6955
All	1094	7906	9000

Conclusiones

Luego de efectuar este análisis podríamos hacer énfasis en los factores que mejor respuesta de pago mostraron:

- En cuanto a género hay una mejor respuesta por parte de las mujeres.
- A nivel educativo, aunque un poco sorpresivo es el nivel escolar que mostró mejor respuesta positiva, seguido del universitario. Antes de hacer el estudio pensábamos que en primer lugar estarían los universitarios al pensar que podrían tener un nivel de ingreso superior.
- Los solteros son los que mejor respuesta a pagar presentaron en cuanto a estado civil.
- En cuanto a rangos de edad, son los menores a 35 años los que mejor proporción de pago muestran.
- A medida que los montos de préstamo suben, la respuesta a pagar el crédito se incrementa. Sin embargo el grueso de préstamos se encuentra hasta los $400mil.

Como se mencionó en el apartado de Modelos, si bien es cierto que el Random Forest del experimento 1 es el mejor modelo de todos los revisados con una exactitud que podría llegar al 85%, este modelo incluye variables como los estados de pago históricos de 6 meses que en el momento de valorar a un cliente son datos desconocidos y por tanto no tiene valor, por esta razón y como lo dijimos, nuestra opción es el modelo KNN con datos conocidos, aunque perdemos en exactitud (79%), es un modelo que trabaja con los datos reales de un cliente potencial.

Por último, podríamos recomendar que en una posible segunda etapa de este proyecto, habrían datos que podrían agregar más valor al estudio, como lo es el ingreso económico, tabular si el cliente tiene otras deudas (casa, carro, etc), miembros en la familia (si la tiene), entre otros.