Nonpayment of credit cards

TWBANK
TAIWAN-BANK

Erick Aguilar
Andrea Quero

CODERHOUSE

Tabla de Contenido

1. Descripción del Caso de Negocio

La banca extiende tarjetas de crédito a un gran número de personas. Al cobrarles intereses y comisiones anuales generan sus ganancias. Sin embargo, si extendieron esas tarjetas a una gran cantidad de personas que no pudieran pagar las deudas adquiridas, entonces el banco podría enfrentar serios problemas de liquidez y gastos inesperados para tratar de recuperar su dinero.

Se desea encontrar el perfil de los usuarios de tarjetas de crédito que tienen más probabilidades de caer en impago (default). Así mismo, se busca encontrar un patrón o tendencia que sugiera que un usuario está próximo a caer en impago.

2. Objetivos del Modelo

El Dataset proporcionado por el banco dispone de registros que contienen información sobre factores demográficos, datos crediticios, historial de pago, estados de cuenta y si han caído o no en impago (default) sus clientes con tarjeta de crédito en TW-Bank entre abril 2005 a septiembre del mismo año.

En función del interés del banco en extender tarjetas de crédito y cobrar los intereses correspondientes, se hace importante evitar extender tarjetas a personas con altas probabilidades de caer en impago de las deudas adquiridas y de esta forma el banco enfrentar problemas de liquidez. Así, TW-Bank está interesado en: a) encontrar los perfiles de las personas que más tienden a caer en impago, con la finalidad de ser más rigurosos con estos al momento de aprobar tarjetas, y b) detectar los patrones/tendencias en las cuentas de las personas que sugieran que caerán en impago en un futuro cercano con la finalidad de tomar medidas previsorias antes de que esto ocurra.

3. Descripción de los Datos

El dataset utilizado corresponde a datos públicos que han sido descargados del sitio web Kaggle.

Las variables originales que comprende dicho dataset se describen a continuación:

- **ID**: Identificador de cada cliente

- **LIMIT_BAL:** Cantidad de crédito otorgado en Nuevo Dolar Taiwanés (incluye tarjeta individual y adicionales)
- **SEX**: Género (1=masculino, 2=femenino)
- **EDUCATION**: Nivel de estudios (1=estudios de posgrado, 2=universidad, 3=preparatoria, 4=otros, 5=desconocido, 6=desconocido)
- **MARRIAGE**: Estado civil (1=casado, 2=soltero, 3=otros)
- **AGE**: Edad en años
- **PAY_0:** Estado de pago en septiembre del 2005 (-1=pago en tiempo y forma, 1=pago retrasado por un mes, 2=pago retrasado por dos meses, … 8=pago retrasado por ocho meses, 9=pago retrasado por nueve meses o más)
- **PAY_2:** Estado de pago en agosto del 2005 (misma escala para PAY_0)
- **PAY_3:** Estado de pago en Julio del 2005 (misma escala para PAY_0)
- **PAY_4:** Estado de pago en junio del 2005 (misma escala para PAY_0)
- **PAY_5:** Estado de pago en mayo del 2005 (misma escala para PAY_0)
- **PAY_6:** Estado de pago en abril del 2005 (misma escala para PAY_0)
- **BILL_AMT1**: Saldo del Estado de Cuenta en septiembre del 2005
- **BILL_AMT2**: Saldo del Estado de Cuenta en agosto del 2005
- **BILL_AMT3:** Saldo del Estado de Cuenta en Julio del 2005
- **BILL_AMT4**: Saldo del Estado de Cuenta en junio del 2005
- **BILL_AMT5**: Saldo del Estado de Cuenta en mayo del 2005
- **BILL_AMT6:** Saldo del Estado de Cuenta en abril del
- **PAY_AMT1:** Cantidad del pago previo en septiembre del 2005
- **PAY_AMT2**: Cantidad del pago previo en agosto del 2005
- **PAY_AMT3:** Cantidad del pago previo en julio del 2005
- **PAY_AMT4:** Cantidad del pago previo en junio del 2005
- **PAY_AMT5:** Cantidad del pago previo en mayo del 2005
- **PAY_AMT6:** Cantidad del pago previo en abril del
- **default.payment.next.month:** Impago en el siguiente mes (1=si, 0=no)

Nota: variables PAY, BILL_AMT y PAY_AMT están expresadas en unidades de nuevo dólar taiwanes

La cantidad del saldo del Estado de Cuenta, la cantidad de Pagos y el límite de crédito varían mucho de persona a persona, por lo que mostraron una amplia dispersión. Por esta razón se decidió combinar estas variables en una sola: La *deuda* expresada como porcentaje del límite de crédito

La *deuda* del mes X se define como el saldo del Estado de Cuenta en el mes X menos el pago hecho en el mes X+1. Esta deuda se divide entre el límite de crédito de cada usuario para expresarlo en términos de porcentaje y así disminuir el efecto de los valores atípicos (outliers).

Se observó que las columnas PAY_n muestran categorías no definidas: "0" y "-2". De acuerdo a la información proporcionada con el dataset, los número positivos reflejan la cantidad de meses de atraso, se asumió que las categorías "0" y "-2" deben corresponder a la categoría "-1", es decir, que se pagó en tiempo y forma, por lo que se realizó esta recategorización de la variable.

A continuación, se muestra una tabla resumen de todas las variables, incluidas las variables porcentaje de déuda (PercOfAprDebt, PercOfMayDebt, PercOfJunDebt, PercOfJulDebt y PercOfAugDebt) para abril, mayo, junio, julio y agosto respectivamente.

Tabla 1 Resumen de variables del Dataset

Columna	tipo	Datos No-Nulos	Datos Nulos	Datos únicos	Ejemplo
ID	int64	30000	0	30000	1
LIMIT_BAL	float64	30000	0	81	20000
SEX	int64	30000	0	2	2
EDUCATION	int64	30000	0	4	2
MARRIAGE	int64	30000	0	3	1
AGE	int64	30000	0	56	24
PAY_0	int64	30000	0	9	2
PAY_2	int64	30000	0	9	2
PAY_3	int64	30000	0	9	-1
PAY_4	int64	30000	0	9	-1
PAY_5	int64	30000	0	8	-1
PAY_6	int64	30000	0	8	-1
BILL_AMT1	float64	30000	0	22722	3913
BILL_AMT2	float64	30000	0	22345	3102
BILL_AMT3	float64	30000	0	22025	689
BILL_AMT4	float64	30000	0	21547	0
BILL_AMT5	float64	30000	0	21009	0

BILL_AMT6	float64	30000	0	20603	0
PAY_AMT1	float64	30000	0	7942	0
PAY_AMT2	float64	30000	0	7898	689
PAY_AMT3	float64	30000	0	7517	0
PAY_AMT4	float64	30000	0	6936	0
PAY_AMT5	float64	30000	0	6896	0
PAY_AMT6	float64	30000	0	6938	0
default.payment.next.month	int64	30000	0	1	1
PercOfAprDebt	float64	30000	0	24873	0
PercOfMayDebt	float64	30000	0	25186	0
PercOfJunDebt	float64	30000	0	25474	0.03445
PercOfJulDebt	float64	30000	0	25729	0.1551
PercOfAugDebt	float64	30000	0	26225	0.1612

De esta manera, el dataset cuenta con 30 variables: 3 variables categóricas y 27 variables numéricas, de las cuales 5 son variables numéricas calculadas para facilitar el análisis de los datos.

4. EDA: Exploratory Data Analysis

Se realizó un análisis exploratorio de los datos, tanto numéricos como categóricos a modo de explorar las siguientes hipótesis:

Hipótesis de Variables demográficas:

- Las mujeres tienden a caer más en impago
- Las personas con un menor nivel de estudios tienden a caer más en impagos
- Las personas casadas tienden a caer más en impago
- Las personas jóvenes tienden a caer más en impago

Al analizar la proporción de hombres y mujeres en el dataset, se observó que el 57% de los clientes son mujeres y el 43% hombres (figura 1 A) de una población total de 30,000 clientes. Analizando cuántas mujeres y hombres cayeron en impago en el periodo de tiempo que comprenden los datos, se observa que hay una mayor incidencia de mujeres en impago, sin embargo, hay que considerar la mayor cantidad de mujeres con respecto a la población total. Por esto, se representa la probabilidad de impago por género representado en porcentaje (figura 1 C). Así, las mujeres tienen un 21% de

probabilidades de caer en impago, mientras que los hombres tienen un 24% de probabilidades.

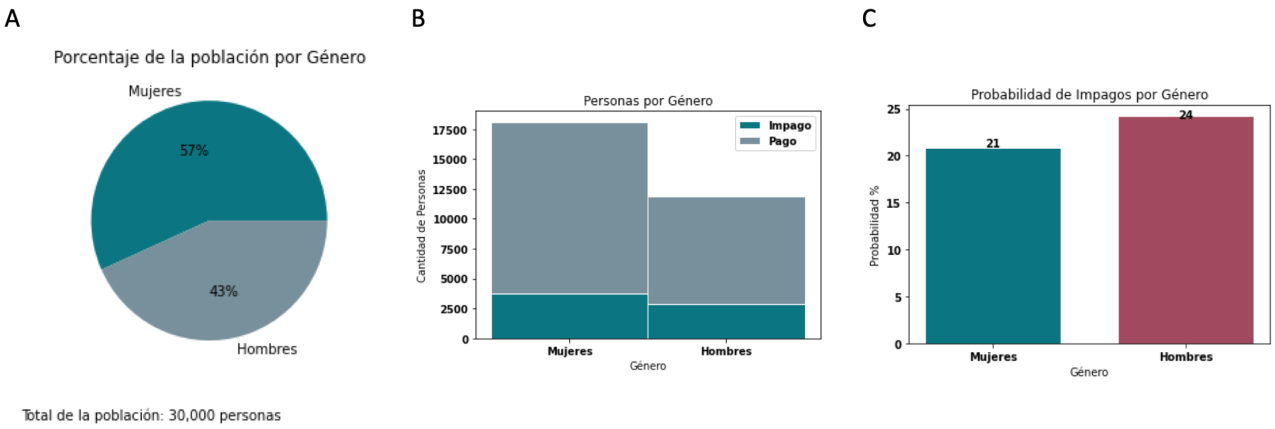

Figura 1 Estudio por género de probabilidad de impago

Posteriormente, se analizó las probabilidades de caer en impago en función del nivel de estudio de los clientes. Primeramente, en la categoría otros se condensaron las categorías *desconocidos* y sin categorizar. Observando la proporción, se observa que la mayoría de los usuarios tienen estudios universitarios, seguidos por los que tienen estudios de posgrado, los que tienen preparatoria y finalmente la categoría otros (Figura 2 A). A partir de estos datos, se calculó la probabilidad de caer en impago en función del nivel de estudio. A pesar de que la proporción de clientes con menor nivel de estudio o un nivel desconocido es menor en cuanto a la disponibilidad de datos en el dataset, se observa que es la categoría con mayor probabilidad de caer en impago, seguido por personas de nivel universitario y finalmente posgrado (Figura 2 B)

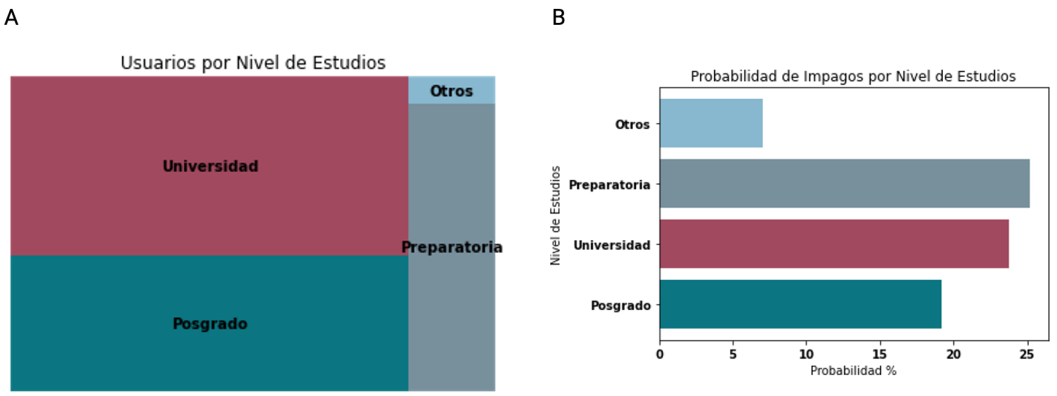

Figura 2 Proporción de usuario en función de nivel de estudio y probabilidad de impago

El dataset muestra una mayor proporción de clientes solteros, seguido de casados y finalmente aquellos designados a la categoría otros, con un 53, 46 y 1 % respectivamente (figura 3 A). Analizando la probabilidad de que los clientes caigan en impago en función del estado civil se encontró que los clientes de la categoría otros tienen mayor posibilidad de caer en impago, pero solo un 1% más de probabilidades que los clientes casados y 3% más probabilidades que los clientes solteros (Figura 3 B). Por lo tanto, esta información por sí sola no es suficiente para predecir si un cliente es propenso a caer en impago o no.

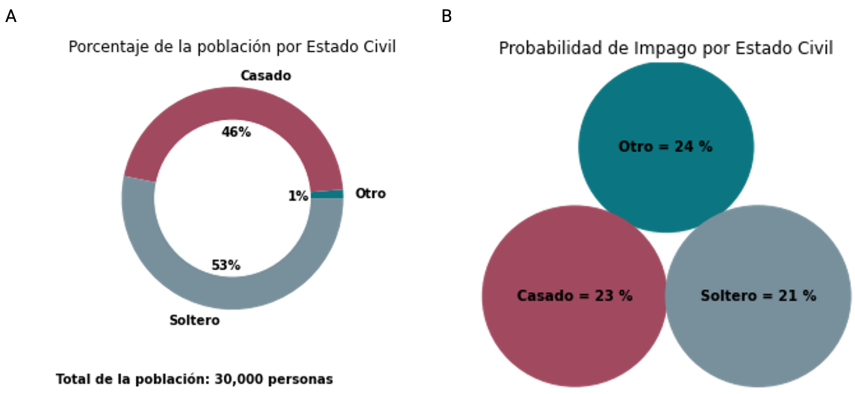

Figura 3 Porcentaje de población por estado civil y su probabilidad de caer en impago

Una variable para considerar que puede afectar el pago de una deuda de una persona es la edad. El dataset muestra una mayor concentración de clientes entre las edades de 20 a 40 años, disminuyendo progresivamente (figura 4 A). Al analizar la cantidad de usuarios que cayeron en impago dentro de este rango de edades se observa que es proporcional a la cantidad de usuarios (figura 4 B), sin embargo, a tomar en cuenta la probabilidad de impago en función de la edad la probabilidad de impago aumenta considerablemente a medida que aumenta la edad del cliente (figura 4 C).

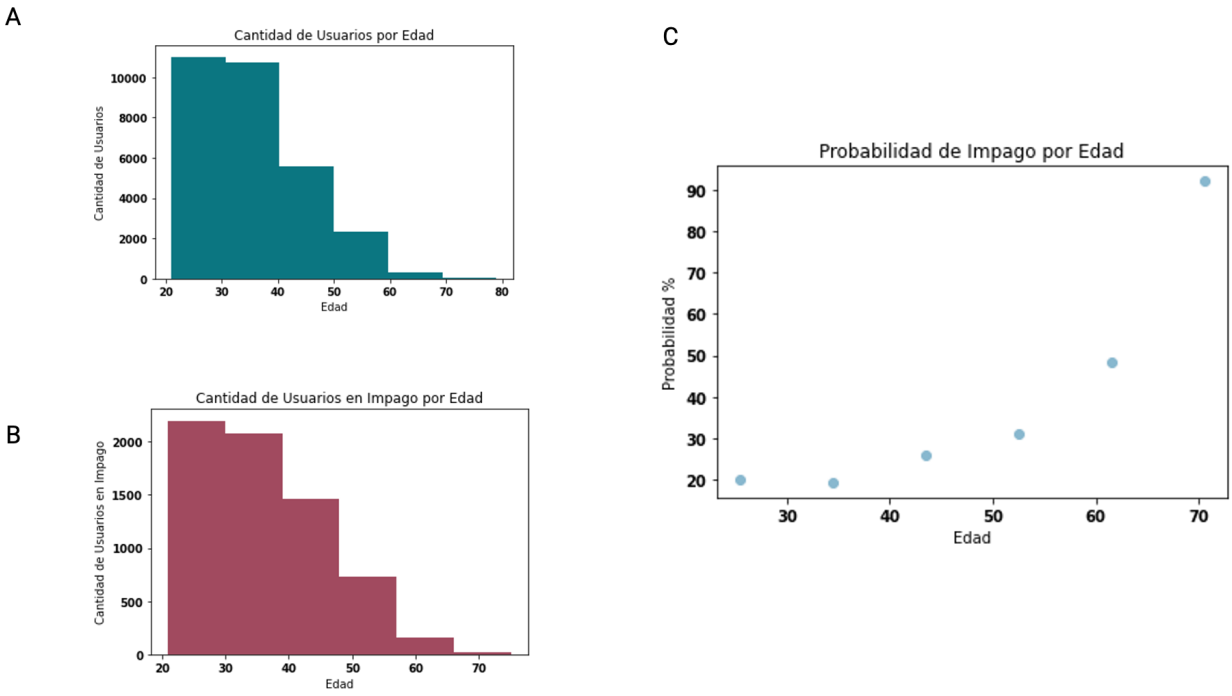

Figura 4 Proporción de clientes por edad y probabilidad de impago

Conclusiones de las variables demográficas:

Las probabilidades de caer impago aumentan

1. Cuando el género del usuario es masculino
2. Conforme el usuario tiene un menor nivel de estudios
3. Cuando el usuario no es soltero
4. Exponencialmente conforme aumenta la edad del usuario

Por lo tanto, se recomienda al banco poner atención especial en el proceso de aprobación de tarjetas de crédito para personas del sexo masculino, con un nivel de estudio bajo, no solteras y de edad avanzada, ya que éstas tienen una probabilidad de caer en impago del 24% y en los registros proporcionados representan un 0.9% de su población.

Hipótesis respecto a las variables históricas y financieras:

- Las personas cuya deuda está más próxima a su límite de crédito tienden a caer más en impago
- Las personas que se han retrasado más en sus pagos tienden a caer más en impago

A medida que una persona cuenta con una deuda más próxima a su límite de crédito, puede ser más complicado saldar la misma de manera que las probabilidades de que aumente la deuda y los clientes caigan en impago son cada vez mayores.

La mediana del saldo de los estados de cuenta aumenta en mayor proporción de lo que aumentan los pagos a través del tiempo. Como se mencionaba anteriormente, la cantidad de Pagos y el límite de crédito varían mucho de persona a persona, generando una amplia dispersión de los datos. Por esta razón se decidió combinar estas variables en una sola: La deuda expresada como porcentaje del límite de crédito. Se procedió a analizar la correlación entre las variables utilizando la correlación de Pearson, recomendada para obtener la correlación entre las variables continuas (*porcentaje de deuda*). De la matriz de correlación se observó que, de las variables *Porcentaje de Deuda* la que tiene más influencia (mayor correlación) con la variable de salida buscada (*defult.payment.next.month*) es el porcentaje de deuda correspondiente al mes de abril, con una correlación de 0.12. Esta correlación es muy baja, por lo que no existe una evidencia estadística fuerte de que las personas cuya deuda está más próxima a su límite de crédito tienden a caer más en impago.

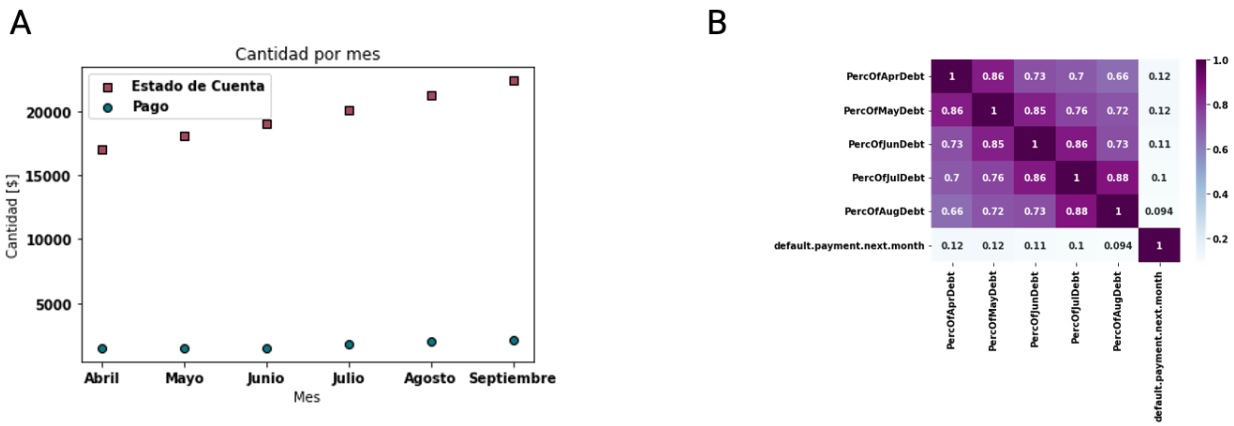

Figura 5 Relación entre impago y cercanía al límite de crédito de la deuda

Analizando la relación entre la salida (condición de impago) y el estado de pago (meses de retraso) en los diferentes meses que comprenden los datos, se observó que la mayor correlación (0.4) se da con el estado de pago en septiembre (figura 6), esto indica que hay una correlación mediana entre ambas variables. Por tal motivo, se encuentra cierta relación entre los meses de retraso en septiembre y el estado de impago del cliente.

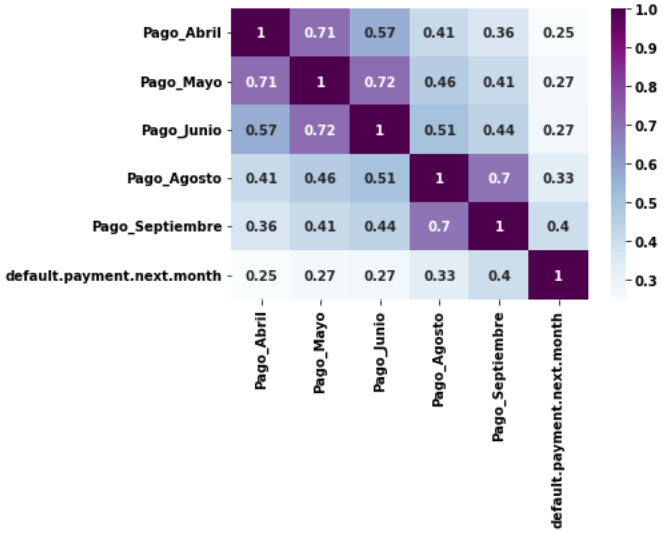

Figura 6 correlación entre retrasos en el pago y probabilidad de caer en impago

Conclusiones de las variables históricas/financieras

1. Se encontró una relación baja entre a) la deuda expresado como porcentaje del límite de crédito y b) la condición de impago
2. Se encontró una relación mediana entre a) los meses de atraso en el pago (en el mes de Septiembre) con la condición de impago

Por lo tanto, se recomienda al banco poner atención detallada en aquellos usuarios que empiezan a atrasarse en su pago, ya que es posible que lleguen a caer en impago.

Conclusiones Generales del Análisis Exploratorio

1. En el proceso de aprobación de tarjetas, se recomienda al banco poner atención especial en las solicitudes de personas del sexo masculino, con un nivel de estudio bajo, no solteras y de edad avanzada, ya que este grupo tiene más probabilidades de caer en impago.
2. En el monitoreo de las cuentas, se recomienda al banco poner atención detallada en aquellos usuarios que empiezan a atrasarse en su pago, ya que es posible que lleguen a caer en impago.
3. En el monitoreo de las cuentas, se sugiere al banco poner atención especial en los usuarios del sexo masculino, con un nivel de estudio bajo, no solteras y de edad avanzada, que empiezan a atrasarse con su pago, ya que el riesgo de que caigan en impago es alto

5. Data Wrangling

Inicialmente, se exploraron los datos a modo de corroborar que estuvieran los datos completos, sin nulos y a qué tipo de datos correspondían, tal como se puedo observar en la tabla 1. Tampoco se encontraron registros duplicados. Las variables saldo y límite de crédito mostraron una amplia

dispersión. Dada la importancia de los datos, los valores atípicos no deben ser eliminados, de manera que fueron condensados en una nueva variable llamada deuda, tal como se explica en la sección 4.

Selección de variables:

La selección de variables para el proceso de entrenamiento del modelo se realizó a través de métodos de filtración, Estadística y Wrapper Methods.

El método ANOVA fue utilizado para identificar cuáles de todas las variables numéricas de entrada tienen mayor influencia en la variable categórica de salida (impago o no impago). Las variables más relevantes, es decir, con un mayor Fs score resultaron las variables 4 a 9 que corresponden a los meses de atraso en pagos para diferentes meses. Las siguientes en relevancia fueron las variables relacionadas a la deuda en los diferentes meses (expresadas en términos de porcentaje del límite de crédito; Figura 7).

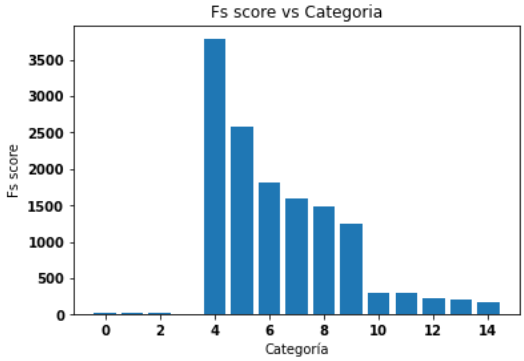

Figura 7 Fs score para cada variable numérica presente en el dataset

Se utilizó una matríz de correlaciones para las variables pago (PAY_N) y porcentaje de deuda (*PercOfAprDebt*). De los análisis ANOVA anteriores, las siguientes variables numéricas seleccionadas para incluirse en el modelo de Machine Learning fueron las siguientes:

1. PAY_0
2. PAY_3
3. PAY_5
4. PercOfAprDebt
5. PercOfAugDebt

Las variables categóricas género, estudio y estado civil fueron seleccionadas mediante el método de chi-cuadrada. Mediante este método se determinó que la categoría de salida (impago) es dependiente de estas tres variables, por lo que son relevantes para el estado de impago del cliente. Así, se seleccionaron para su inclusión en el modelo de Machine Learning.

One Hot Encoding:

Antes de proceder con el proceso de modelado de Machine Learning se realizó un One Hot Encoding para las variables categóricas género, estudio y estado civil; de esta manera, evitar que el algoritmo asigne mayor peso a una categoría que a otra de manera incorrecta.

Preparación de subconjuntos de entrenamiento y prueba:

Durante el análisis exploratorio de los datos se observó que la probabilidad de que un usuario caiga en impago es del 22.12%, lo que significa que en el set de datos originales está desbalanceado con respecto a la variable de salida que se deseaba predecir (impago).

Para compensar el desbalance de la variable y disminuir la posibilidad de un overfitting, se decidió realizar un **oversampling**: Aumentar artificialmente (repitiendo instancias) las instancias positivas, es decir, la variable impago. Esta decisión se sustenta en que los modelos de Machine Learning aprenden mejor a medida que se les muestran más instancias y esta técnica da la flexibilidad de ocupar varios modelos de Machine Learning y no sólo limitarlo a árboles de decisión.

Escalamiento:

Por último, se procedió al escalmiento de los datos. El escalar significa dividir los datos entre su desviación estándar. Notar que el escalado de datos se realiza después de haber dividido los datos para prueba y entrenamiento con la finalidad de evitar fuga de datos. La fuga de datos ocurre cuando la información acerca del dataset para entrenamiento corrompe o influye en el dataset de prueba.

6. Modelos Implementados

El problema presentado corresponde a un problema de clasificación. Debido a que el set de datos utilizado muestra si el cliente cayó o no en impago, nos referimos a un problema de clasificación supervisado, más exactamente.

El objetivo de este proyecto es detectar cuando un cliente caerá en impago para evitarle pérdidas al banco, por lo que se ocupó la exhaustividad o recall para evaluar el desempeño de los modelos obtenidos.

Árbol de decisión:

Inicialmente se realizó el árbol de decisión sin incluir la variable Edad, y posteriormente utilizando esta variable dado que anteriormente se observó que la probabilidad de caer en impago aumenta exponencialmente con la edad.

Tabla 2 Resultados para la predicción de la variable impago sin incluir la variable edad con árbol de decisión.

	Exhaustividad para el caso positivo	Exactitud
test	0.95	0.86
Train	1.00	0.98

El árbol de decisión creado tuvo muy buena exhaustividad para predecir los casos de impago del dataset de prueba (95%). Sin embargo, el que sea capaz de predecirlos al 100% con el dataset de entrenamiento deja la impresión de que este modelo está sobreajustado. En ambos casos, a modo

de quitar este sobreajuste, se procedió a buscar un valor óptimo para el parámetro Alpha para con él, "podar" del árbol.

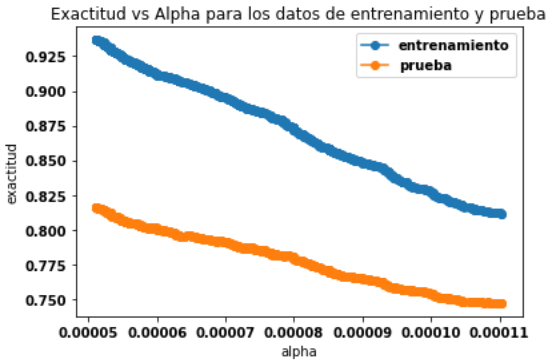

Figura 8 Exactitud de los árboles en función de Alpha para el set de datos de entrenamiento y prueba.

En la figura 8 se muestra que conforme aumenta el valor de Alpha, el valor de la exactitud disminuye, tanto para el set de datos de entrenamiento como de prueba; y aunque no se interceptan, se acercan entre sí. Se decidió que un alpha de 0.00006 proporciona una exactitud de alrededor del 80% entre los datos de entrenamiento y de prueba. Esta exactitud similar asegura que no se tenga un sobreajuste con el dataset de entrenamiento.

Tabla 3 Resultados para la predicción de la variable impago sin incluir la variable edad con árbol de decisión. Alpha= 0.00008

	Exhaustividad para el caso positivo de impago	Exactitud
test	0.88	0.81
Train	0.95	0.91

Aunque el árbol con el ajuste en el valor de Alpha tiene una exhaustividad menor con los datos de prueba comparado con el árbol sin podar (81% vs 95%) para el caso de impago, no tiene una exhaustividad perfecta con los datos de entrenamiento y la exactitud de ambos no es tan dispar (81% vs 91%), por lo que se concluyó que este árbol no está sobreajustado y pudiera tener un mejor desempeño ante nuevos datos.

Al incluir la variable edad, que como se mencionaba anteriormente, la probabilidad de caer en impago aumenta exponencialmente con la edad, se observa lo siguiente:

Tabla 4 Resultados para la predicción de la variable impago incluyendo la variable edad con árbol de decisión.

	Exhaustividad para el caso positivo de impago	Exactitud
test	0.95	0.85
Train	1.00	0.99

A pesar de que la exhaustividad para predecir impagos (recall) es del 95%, el modelo da la impresión de estar sobreajustado ya que el desempeño utilizando los datos de entrenamiento es del 100%.

Se procedió a ajustar el Alpha utilizando la gráfica de exactitud vs el valor de Alpha para los set de datos de entrenamiento y prueba como se muestra en la figura 9.

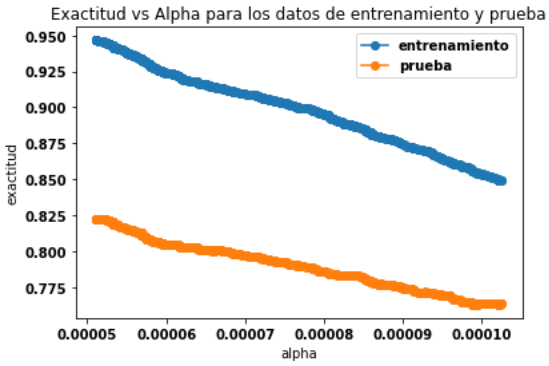

Figura 9 Exactitud de los árboles en función de Alpha para el set de datos de entrenamiento y prueba incluyendo la variable edad.

Se decide que con un Alpha de 0.000085 se tiene una exactitud del 80% aproximadamente para ambos datasets, a modo de asegurar que no exista un sobreajuste del modelo.

Tabla 5 Resultados para la predicción de la variable impago incluyendo la variable edad con árbol de decisión. Alpha= 0.000085.

	Exhaustividad para el caso positivo de impago	Exactitud
test	0.84	0.78
Train	0.92	0.89

A pesar de que el árbol al que se le realizó el ajuste del hiperparámetro Alpha muestra una menor exhaustividad (recall) que el árbol original, considerando en ambos casos la variable edad; la exactitud entre los datos de prueba y entrenamiento luego de ajustar el valor de Alpha no son muy diferentes entre sí y distintos del 100%. Se concluyó que el árbol luego de la modificación del valor de Alpha no está sobre ajustado y, por lo tanto, pudiera tener un mejor desempeño ante nuevos datos.

Comparación del árbol con edad y sin edad:

Al comparar la exhaustividad obtenida del árbol podado sin la variable edad (88%) con la obtenida del árbol podado que incluye la variable edad (84%), se concluye que la variable edad no tiene una influencia fuerte en el resultado (clasificación como pago o impago). Esto concuerda con los resultados obtenidos a través de los métodos wrapping y estadísticos para la selección de variables a incluir en el modelo.

Del mismo modo, el árbol de decisión va seleccionando las variables que tienen más influencia para la clasificación de los resultados. En el árbol creado tomando en cuenta la variable edad, se aprecia que el comportamiento de pagos (PAY, PercofDeb) tiene un peso mucho mayor que las variables demográficas (AGE, SEX, etc), ya que PAY y PercofDeb aparecen en los primeros nodos.

Ramdom Forest:

Inicialmente se realizó una primera aproximación con Random Forest utilizando los hiperparámetros que éste modelo da por defecto.

Tabla 6 Resultados para la predicción de la variable impago mediante Random Forest

	Exhaustividad para el caso positivo de impago	Exactitud
test	0.95	0.87
Train	1.0	0.98

Aunque este bosque aleatorio tiene una muy buena exhaustividad para predecir impagos (95%) se presume que está sobreajustado, ya que su desempeño con el dataset de entrenamiento es de 100%

Se realizó un ajuste de hiperparámetros utilizando **RandomizedSearchCV**. Se determinó que los parámetros óptimos son n_estimators: 4, max_feature='log2', max_depth: 37, criterion = 'entropy'. A partir de estos hioperparámetros se obtuvieron los siguientes resultados:

Tabla 7 Resultados para la predicción de la variable impago mediante Random Forest optimizado con RandomizedSearchCV

	Exhaustividad para la condición positiva de impago	Exactitud
test	0.89	0.86
Train	0.97	0.97

Se realizó un tercer ajuste de hiperparámetros utilizando **GridSearchCV** Con la finalidad de hacer un ajuste fino de los mismos. Se determinó que los parámetros óptimos son criterion = 'entropy', max_depth = 30, max_features = 'auto', y n_estimators: 4. A partir de estos hioperparámetros se obtuvieron los siguientes resultados:

Tabla 8 Resultados para la predicción de la variable impago mediante Random Forest optimizado con GridSearchCV

	Exhaustividad para la condición positiva de impago	Exactitud
test	0.88	0.85
Train	0.97	0.96

El bosque aleatorio encontrado, aunque tiene una exhaustividad menor con los datos de prueba, comparado con el bosque sin ajuste de hiperparámetros (88% vs 95%) para el caso positivo de impago, no tiene una exhaustividad perfecta con los datos de entrenamiento y sus exactitudes no son tan diferentes (85% vs 96%), por lo que se concluye que este árbol no está sobreajustado y pudiera tener un mejor desempeño ante nuevos datos.

Random Forest - Comparación:

Se puede observar que la exhaustividad obtenida en la primera aproximación fue la mejor, ya que posteriormente los hiperparámetros fueron ajustados con ayuda de la función RandomizedSearchCV con la finalidad de obtener un modelo no sobreajustado. Del mismo modo, se puede observar que esta exhaustividad prácticamente no cambió al intentar hacer una búsqueda más refinada con ayuda de la función GridSearchCV. De lo anterior, se concluye que una búsqueda aleatoria (RandomizedSearchCV), dependiendo del número de muestras que tome, puede ayudar a encontrar hiperparámetros mejorados con un costo computacional menor comparado con el requerido por GridSearchCV.

Support Vector Machine con Análisis de Componentes Principales

Para realizar un análisis de componentes principales o PCA, por sus siglas en inglés, fué necesario tener los datos centrados y en la misma escala. El centrar los datos significa hacer que su media sea igual a cero y escalado quiere decir dividir los datos entre su desviación estándar.

Antes de aplicar PCA teníamos 15 variables. PCA las redujo a 12, como puede apreciarse en la Figura 10 (12 barras). En la misma figura se observa que el componente principal 1 (PC1) es responsable de alrededor del 17% de la variación, seguido por el PC2 al cual se le puede atribuir alrededor del 16%, el PC3 alrededor del 14%, el PC4 alrededor del 11% y así sucesivamente.

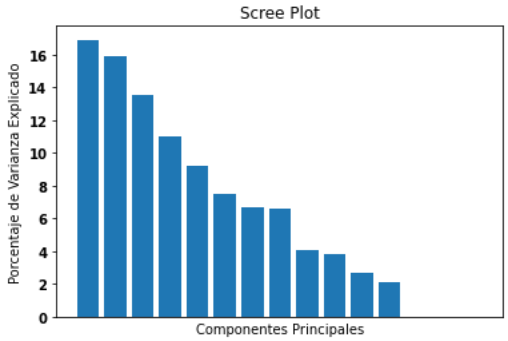

Figura 10 Gráfica de componente principales presentes en el Dataset

Dado que la reducción de variables no fue sustancial (de 15 a 12), se decidió ocupar únicamente los componentes principales que tuvieran un efecto en la variabilidad mayor o igual al 10%. Por lo tanto, para el entrenamiento del SVM, únicamente se tomaron PC1, PC2, PC3 y PC4. Los hiperparámetros óptimos se obtuvieron utilizando RandomizedSearchCV resultando ser C: =100, gamma = 1 y kernel = 'rbf'. Utilizando estos hiperparámetros se procede a ejecutar el Suport Vector Machine (SVM) obteniendo los siguientes resultados:

	Exhaustividad para el caso positivo de impago	Exactitud
test	0.54	0.42

Train	0.62	0.71

Con los datos de entrenamiento, tomando como referencia la exhaustividad para el caso positivo de impago, el desempeño mejora sólo un poco (de 54% a 62%). Dado que el desempeño es relativamente pobre en ambos casos, se concluye que el modelo está subajustado.

Este subajuste se atribuye a que los componentes principales utilizados para el entrenamiento del modelo, en su conjunto sólo son responsables de alrededor de un 58% de la variabilidad (PC1 17% + PC2 16% + PC3 14% + PC4 11%). Es decir, en este dataset no hay componentes principales que tengan un peso mucho mayor en comparación de otros con los cuáles se pueda explicar un 80% o más de la variabilidad.

Aunque este modelo no parece muy prometedor para el objetivo del proyecto, tomaremos ventaja de él a través de la Figura 11 para conocer cuáles son las variables que considera con más influencia sobre la Componente Principal 1 (responsable de la mayor variación).

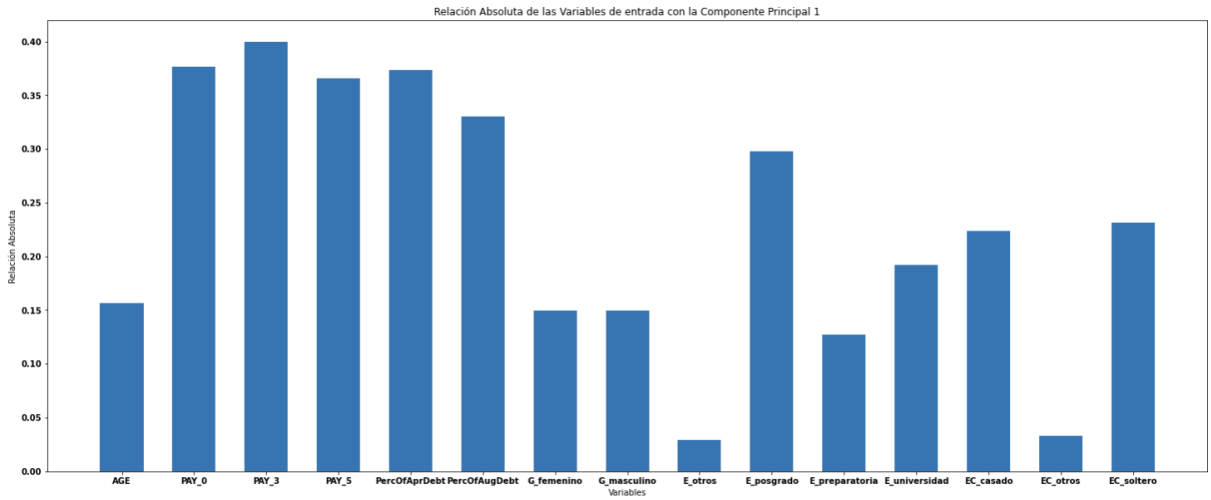

Figura 11 Relación Absoluta de las Variables de Entrada con la Componente Principal 1

En la gráfica anterior puede apreciarse que las variables con mayor influencia son las relacionadas al comportamiento de pago (PAY,

PercOfDebt) y no las variables demográficas como nivel de estudio, estado civil, edad, etc.

7. Conclusiones

1. Se encontró que las variables históricas y financieras tienen una mayor influencia sobre la salida (condición de pago o impago) que las variables demográficas. Es decir, que el comportamiento de los clientes respecto a qué tanto se endeudan y qué tan bien han pagado sus deudas a través del tiempo es un mejor predictor que las variables demográficas respecto a si caerán o no en una condición de impago.
2. Al balancear (oversampling) la variable de salida para obtener un número similar en las instancias tanto positivas como negativas de impago, se observó una mejora tanto en el recall como en la precisión de los casos positivos.
3. Se logró obtener un buen modelo predictivo a través de un árbol de decisión podado, con una exhaustividad (recall) del 84%.
4. Dicho modelo se pudo mejorar al usar un bosque en lugar de un sólo árbol. Con un bosque la exhaustividad logro aumentarse del 84% al 88%.
5. El modelo creado usando Support Vector Machine con Análisis de Componentes Principales no resultó ser muy útil para este problema. Esto es debido a que la variabilidad no se concentra sólo en algunos de los Componentes Principales encontrados, sino que se distribuye en varios de ellos, por lo que la reducción de dimensionalidad compromete bastante el desempeño del modelo.
6. Support Vector Machine con Análisis de Componentes Principales fue el modelo que más recursos computacionales requirió, pero que, a la vez, el que demostró el menor desempeño. Por lo que queda demostrado una vez más que un mayor gasto computacional no necesariamente se va a traducir en un mejor desempeño.
7. Si se deseará mandar a producción este proyecto, el modelo seleccionado sería el obtenido a través de Random Forest, el cual

demostró un mejor desempeño en la exhaustividad para el caso de impago. De acuerdo a los datos obtenidos, este modelo sería capaz de predecir el 88% de los casos que caerán en impago.

Futuras mejoras posibles:

Se decidió no utilizar las primeras aproximaciones tanto del Árbol de Decisión como del Random Forest ya que ambas muestran una exhaustividad (recall) del 100% con los datos de entrenamiento para el caso positivo de impago, lo cual puede interpretarse como que dichos modelos están sobreajustados.

Sin embargo, al seleccionar una medida de desempeño un poco más genérica como lo es la exactitud (accuracy), dicha medida muestra un 98% con los datos de entrenamiento y un 87% con los datos de prueba. Se considera que el modelo está sobreajustado cuando se obtiene un muy buen desempeño con los datos de entrenamiento, pero presenta un desempeño pobre con los datos nuevos o de prueba. Este 87%, aunque es un poco menor que el 98%, realmente no es un desempeño pobre. Por lo tanto, sería bueno evaluar los modelos obtenidos en las primeras aproximaciones con más datos nuevos conforme se vayan recopilando antes de descartarlos por completo. Estos modelos, en teoría, tienen un mejor desempeño que los modelos obtenidos ajustando los hiperparámetros y sería bueno evaluarlos de manera más exhaustiva con más datos nuevos antes de desecharlos.