Leonela Tacure P., Víctor Villacorta P., Oscar Rafael T.

Posgrado Ciencia de Datos, Universidad Ricardo Palma, Lima

[Leonela.tp@gmail.com](mailto:Leonela.tp@gmail.com), [v.villacorta.unmsm@gmai.com](mailto:v.villacorta.unmsm@gmai.com), [orafaelp@alicorp.com](mailto:orafaelp@alicorp.com)

Minería de datos para determinar incumplimiento de pagos en créditos para Banco Colombia

Resumen –. El sistema financiero (SF) en Colombia, requiere implementar las nuevas tendencias en minería de datos para tener una ventaja competitiva frente a los demás Bancos, utilizando sus datos y conocimientos como principal activo. El objetivo de esta investigación es determinar el incumplimiento de pago “Default” en los créditos brindados por Bancolombia. En este sentido, se realizó una prueba psicométrica escrita a los clientes con créditos pendientes de pago para ver como estas influyen en el comportamiento del pago crediticio.

A partir de esto, se realiza un análisis de los datos para determinar la influencia de la prueba psicométrica en el incumplimiento de pago. Posteriormente, se plantean la comparación entre 2 modelos: Regresión Logística y Árboles de Decisión y se contrasta con los indicadores de clase: recall, efectividad y Accuracy, para encontrar el mejor modelo que beneficia a la predicción de incumplimiento de pago.

**Palabras claves – Minería de datos, Incumplimiento, Calificación crediticia, Psicometría, Modelos de clasificación.**

Abstrac -. The financial system (FS) in Colombia requires the implementation of new trends in data mining to have a competitive advantage over other Banks, using their knowledge and data as the main asset. The objective of this investigation is to determine the default payment "Default" in the loans provided by Bancolombia. In this sense, a written personality test was carried out on clients with outstanding loans to see how they influence the behavior of credit payments.

From this, an analysis of the data is carried out to determine the influence of the psychometric test on the default of payment. Subsequently, the comparison between 2 models is proposed: Logistic Regression and Decision Trees and contrast them with the class indicators: recall, effectiveness and Accuracy, to find the best model that benefits the prediction of Default.

**Keywords – Data Mining, Default, Credit Scoring, Classification models.**

# INTRODUCCIÓN

Debido a la crisis financiera de 1998, la Superintendencia Financiera de Colombia (SFC), solicitó a todas las entidades bancarias emplear modelos estadísticos que permitan determinar que un cliente no cumpla con el pago de sus obligaciones. Según la Circular Externa 011 de 2002 de la [Superintendencia Financiera de Colombia (SFC)](https://www.redalyc.org/jatsRepo/2821/282154593002/html/index.html#redalyc_282154593002_ref28), se define “la posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que sus deudores fallen en el cumplimiento oportuno o cumplan imperfectamente los contratos de crédito”.

En este sentido, el objetivo general es proponer una metodología para estimar el riesgo de incumplimiento de pago que los clientes que tienen obligaciones con Banco Colombia. El desarrollo del proyecto de investigación está enfocado hacia el objetivo específico de identificar el mejor modelo para estimar los incumplimientos de pago. INV

ESTIGACIÓ EER\*

# MARCO TEÓRICO

De acuerdo con Atiya (2001) la predicción del incumplimiento de pago ha sido un fenómeno ampliamente estudiado debido a que las entidades necesitan predecir y provisionar sus recursos con la finalidad de evitar caer en bancarrota y maximizar utilidades.

BID Invest (2018), explica el potencial de nuevos enfoques alternativos que evalúan la personalidad, la inteligencia y el carácter de quienes sin historial crediticio se acercan a los bancos a solicitar préstamos, con el fin de determinar si son confiables a la hora de pagar. Klinger, Khwaja y del Carpio (2013) sobre 1.580 PYMES en Perú, Kenia, Colombia y Sudáfrica encontró que es la personalidad del individuo la que está vinculada al patrón de comportamiento de pago.

Ponce Cruz (2010), menciona en su libro que la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan como una función de variables explicativas, utilizando una función logística.

# METODOLOGÍA

## DEFINICIONES

* **Coeficiente de correlación p de Sperman** [12]**:** Rho de Spearman es una medida de dependencia no paramétrica en la cuál se calcula la jerarquía media de las observaciones, dando la relación de dependencia entre dos variables continuas.
* **Matriz de confusión:** Contiene el número de casos bien clasificados y mal clasificados. Los casos bien clasificados se ubican en la diagonal, los elementos de está representan casos mal clasificados.

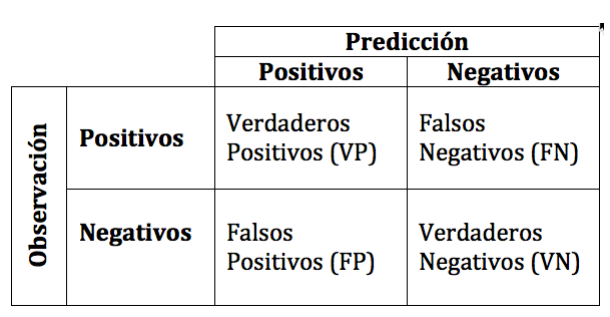


Fig. 1: Distribución matriz de confusión. Zelada Carlos

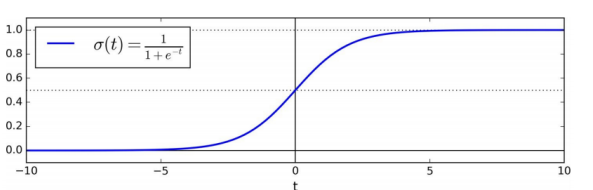
* **Recall:** Mide el porcentaje de los clientes “Incumplidos” somos capaces de identificar.
* **Efectividad:** Mide el porcentaje de clientes “Incumplidos”.
* **Accuracy:** Mide el porcentaje de casos que el modelo ha acertado.
* **Validación cruzada:** Técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la participación entre datos de entrenamiento y prueba.
* **Regresión Logística** [13]**:** Por lo general es utilizada para estimar la probabilidad de que una instancia corresponda a una categoría específica. Si la probabilidad estimada es mayor al 50%, entonces el modelo predice a esa categoría (etiquetada como “1”), o bien predice que no (etiquetada como “0”). Esto lo convierte en un clasificador binario.

Fig. 2: Función Logística. Gerón (2017)

* **Arboles de decisión** [14]: Permiten representar una serie de reglas de forma gráfica con el fin de tomar una decisión sobre la asignación de un valor de salida. Se encuentran compuestos por nodos, ramas y hojas o nodos hoja, cuyas características son:
  + Los nodos son los datos de entrada.
  + Las ramas son grupos de registro en las variables de entrada.
  + Los nodos hojas o las hojas son los valores de la variable de salida.

## DESCRIPCIÓN DEL PROCEDIMIENTO DE ANÁLISIS

Los datos fueron analizados utilizando técnicas de aprendizaje supervisado de clasificación. Para implementar el modelo de aprendizaje supervisado de clasificación se utilizó la metodología CRISP-DM:

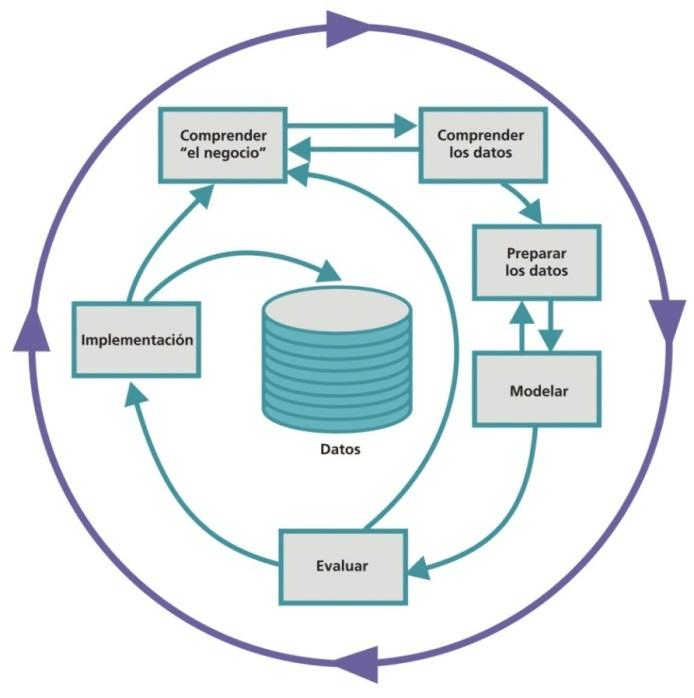


Fig. 3: Fases Metodología CRIP-DM

En la comprensión de los datos, se verifico que cumplía los supuestos para aplicar Regresión Logística, se encontró que las variables predictoras tenían una correlación menor al 0.01 por lo que se realizó ingeniería de variables para encontrar nuevas. Dado que se quiere encontrar cuanto aportan las variables predictoras al Target no se tratan los valores atípicos para no perder información.

Para la fase de preprocesamiento de datos no fue necesario aplicar técnicas de imputación, debido a que nuestros datos no carecían de información en las variables predictoras. Se realizó el reescalamiento de los datos mediante la escala Min-Max. También se realizó feature engineering de acuerdo al peso de nuestras variables mediante combinaciones polinómicas y divisiones aprovechando el tipo continuo de las variables.

Siguiendo el análisis, se utilizó un 85% de los datos para el entrenamiento de los algoritmos de aprendizaje supervisado y 15% para la validación de la performance predictiva de los modelos. Luego, se realizó el entrenamiento del modelo, para esto se utilizó los siguientes experimentos:

* Experimento 1: Se trabajó con las variables predictoras originales.
* Experimento 2: Se trabajó con las variables originales que son significativas a un nivel de confianza mayor a 95%.
* Experimento 3: Se trabajó con los 59 atributos obtenidos mediante ingeniería de características.
* Experimento 4: Se trabajó con los 8 atributos que son significativas a un nivel de confianza mayor a 95%.

Para los experimentos antes mencionados se aplicará 8 técnicas en base a los Modelos de Regresión Logística y Árboles de Decisión, descritas en la TABLA I.

TABLA I

TÉCNICAS UTILIZADAS EN EXPERIMENTOS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cod | Modelo | Transformación | Balanceo | Validación Cruzada | | |
| estra | n-folds | repeat |
| a | Reg. Log | Proporcion min/max | No | No | 0 | 0 |
| b | Reg. Log | Proporcion min/max | No | Si | 3 | 2 |
| c | Reg. Log | Proporcion min/max | Over | Si | 3 | 1 |
| d | Reg. Log | Proporcion min/max | Under | Si | 3 | 1 |
| e | Arb. Desición | ninguno | No | Si | 3 | 1 |
| f | Arb. Desición | ninguno | No | Si | 3 | 2 |
| g | Arb. Desición | ninguno | Over | Si | 3 | 1 |
| h | Arb. Desición | ninguno | Under | Si | 3 | 1 |

Fuente: Elaboración Propia

De los coeficientes generados en los modelos del dataset de entrenamiento se realiza la predicción en el datase de prueba, el punto de corte para la clasificación de un cliente con default se realiza con el índice de Youden. Tomando 0 para “Cumplimiento y 1 para “Incumplimiento”.

Finalmente, para la evaluación del modelo se utilizó el análisis mediante indicadores de clase como son recall, efectividad y accuracy. A continuación, se comparó los indicadores de cada modelo y se eligió el modelo que mejores valores de las métricas obtuvieron para la predicción del incumplimiento de pago.

# PROCEDIMIENTO DESARROLLO DEL TRABAJO

## Recolección de fuentes de información

1. **Fuente de información:** Se cuenta con datos de 80 mil clientes del Banco Colombia que cuenta con préstamos crediticios pendientes.

TABLA II

Fuentes de información

|  |  |
| --- | --- |
| Nombre | Descripción |
| score\_train | Información crediticia de 63’901 clientes con prueba escrita de personalidad. |
| score\_test | Información crediticia de 16’116 clientes con prueba escrita de personalidad. |

Fuente: Elaboración Propia

**2. Operacionalización de variables:**

Describimos a las variables por cada fuente de información

TABLA III

DESCRIPCIÓN BASE SCORE\_TRAIN / SCORE\_TEST

|  |  |
| --- | --- |
| Nombre | Descripción |
| Empleado | 0=Desempleado, 1=Empleado actualmente |
| Tiempo\_empleado | Número de años que lleva en el empleo actual (si aplica) |
| Saldo\_cuenta | Saldo en miles de pesos de la cuenta principal con el banco |
| Valor\_prestamo | Valor actual que adeuda en préstamos en el sistema financiero |
| Cuentas\_otros | Número de cuentas que tiene en otros bancos |
| Autocontrol | Mediante una prueba escrita se mide el nivel de autocontrol de la persona, en puntaje de 0 a 100 |
| Impulsividad | Mediante una prueba escrita se mide el nivel de impulsividad de la persona, en puntaje de 0 a 100 |
| Confianza | Mediante una prueba escrita se mide el nivel de confianza en sí misma de la persona, en puntaje de 0 a 100 |
| Incumplimiento | 0=Cumplimiento; 1=Incumplimiento |

Fuente: Elaboración Propia

Las variables presentan las siguientes características:

* Dicotómicas: (1) Empleado
* Numéricas discretas: (5) Tiempo\_empleo, Cuentas\_otros, Autocontrol, Impulsividad, Confianza.
* Numéricas continuas: (2) Saldo\_cuenta, Valor\_prestamo
* Target: (1) Incumplimiento.

## Análisis exploratorio de los datos

1. **Estadística descriptiva de variables:**

**Target:** clasificación del incumplimiento de pago, está etiquetada como 0 (Cumplimiento) y 1 (Incumplimiento), con un 85.9% y 14.1% respectivamente. Se observa que tenemos un dataset desbalanceado.

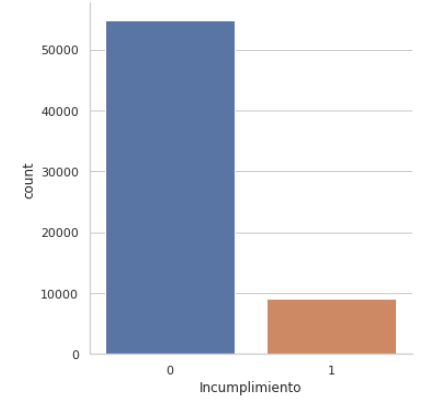


Fig.4 Distribución de variable objetivo

Fuente: Elaboración Propia

**Variables score\_train:** Las variables del dataset son todas variables numéricas, analizando la distribución:

* Las variables de prueba de personalidad aparentemente presentan una distribución normal.
* Las variables de tiempo de empleo y saldo en cuenta, tienen una distribución asimétrica negativa.
* La variable valor préstamo, aparentemente tiene una distribución normal, la cual se debe comprobar con la prueba de normalidad.
* La variable de cuentas en otros bancos no se puede definir si tiene distribución normal o no.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

Fig. 5 Distribución de las variables train

Fuente: Elaboración Propia

## Pruebas estadísticas

1. **Normalidad**

Mediante el Test de D’Agostino obtenemos que las pruebas de personalidad: Impulsividad, Confianza y Autocontrol, evidencian una distribución normal. Rechazando la normalidad de las demás variables.

1. **Homocedasticidad**

Para las variables con distribución Normal se aplica el Test de Bartlet, que nos dice que las pruebas de personalidad: Impulsividad, Confianza y Autocontrol, tienen una varianza constante (homogénea).

Para las variables con distribución No Normal se aplica el Test de Levene, el cual nos dice que las variables tienen una varianza constante (homogénea).

1. **Multicolinealidad**

Mediante el método de Sperman se encontró que las variables no están correlacionadas entre sí, Fig. 6, arrojando correlaciones inferiores al 0.01 respecto al Target, TABLA IV. A partir de estos supuestos se crean unas cuantas variables para testearlas en un modelo de Regresión Logística.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Emple** | **T. empleo** | **Saldo cuenta** | **Valor prest** | **Cuent. otros** | **Auto control** | **Impulsi** | **Confi** | **Target** |
| **Emple** | 1,0000 | 0,6140 | 0,0025 | -0,0055 | -0,0012 | 0,0056 | 0,0053 | 0,0026 | 0,0007 |
| **Tiempo empleo** | 0,6140 | 1,0000 | 0,0025 | -0,0016 | 0,0048 | 0,0029 | -0,0060 | 0,0050 | 0,0010 |
| **Saldo cuenta** | 0,0025 | 0,0025 | 1,0000 | 0,0033 | 0,0006 | -0,0008 | 0,0021 | -0,0066 | -0,0079 |
| **Valor prestamo** | -0,0055 | -0,0016 | 0,0033 | 1,0000 | 0,0002 | 0,0095 | -0,0022 | 0,0009 | -0,0044 |
| **Cuentas otros** | -0,0012 | 0,0048 | 0,0006 | 0,0002 | 1,0000 | -0,0007 | 0,0040 | -0,0010 | 0,0028 |
| **Autocon** | 0,0056 | 0,0029 | -0,0008 | 0,0095 | -0,0007 | 1,0000 | 0,0042 | 0,0076 | -0,0041 |
| **Impuls** | 0,0053 | -0,0060 | 0,0021 | -0,0022 | 0,0040 | 0,0042 | 1,0000 | -0,0038 | -0,0022 |
| **Confi** | 0,0026 | 0,0050 | -0,0066 | 0,0009 | -0,0010 | 0,0076 | -0,0038 | 1,0000 | 0,0004 |
| **Target** | 0,0007 | 0,0010 | -0,0079 | -0,0044 | 0,0028 | -0,0041 | -0,0022 | 0,0004 | 1,0000 |

Fig. 6. Matriz de correlación de variables

Fuente: Elaboración Propia

TABLA IV

correlación de variables predictoras respecto al target

|  |  |
| --- | --- |
| Nombre | TARGET |
| Saldo\_cuenta | 0.00791 |
| Valor\_prestamo | 0.00436 |
| Autocontrol | 0.00407 |
| Cuentas\_otros | 0.00278 |
| Impulsividad | 0.00219 |
| Tiempo\_empleo | 0.00096 |
| Empleado | 0.00068 |
| Confianza | 0.00042 |

Fuente: Elaboración Propia

## Tratamiento de variables

1. En primera instancia se contrastan las variables independientes con la variable target “Incumplimiento”.

Se realiza un escalamiento de las variables por normalización mediante la escala Min-Max. Para luego identificar las variables que ayudarán a predecir el incumplimiento de pago, mediante la identificación de la significancia de las variables, obteniendo Confianza y Autocontrol con mayor importancia de acuerdo al peso y depurando las variables que no aportan al modelo, las que se pueden visualizar en la TABLA V.

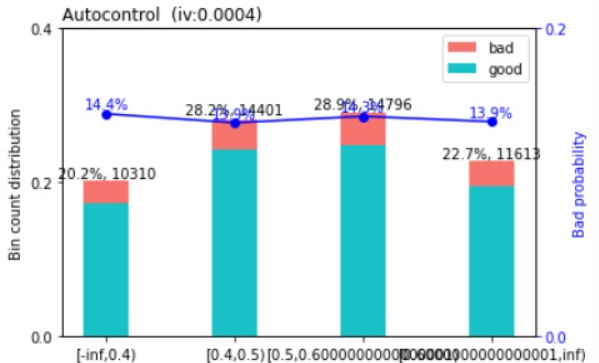
TABLA V

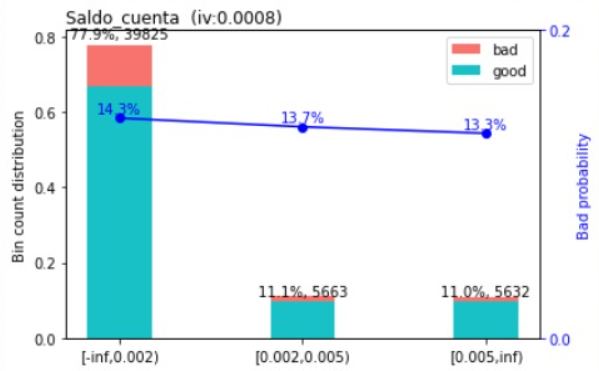
PESOS DE VARIABLES

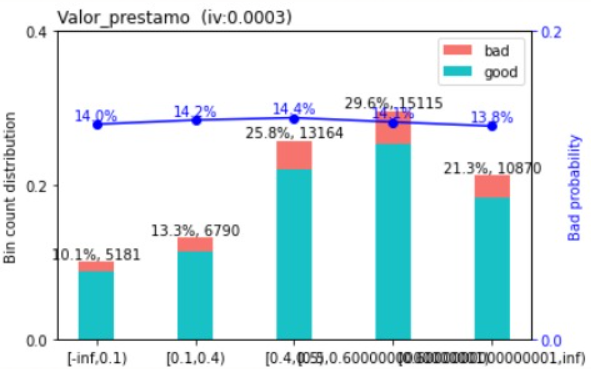
|  |  |
| --- | --- |
| Nombre | Descripción |
| Tiempo\_empleo | 0.047774 |
| Cuentas\_otros | 0.309007 |
| Saldo\_cuenta | 0.993621 |
| Empleado | 5.879682 |
| Valor\_prestamo | 7.954889 |
| Impulsividad | 24.826648 |
| Confianza | 28.326063 |
| Autocontrol | 31.662316 |

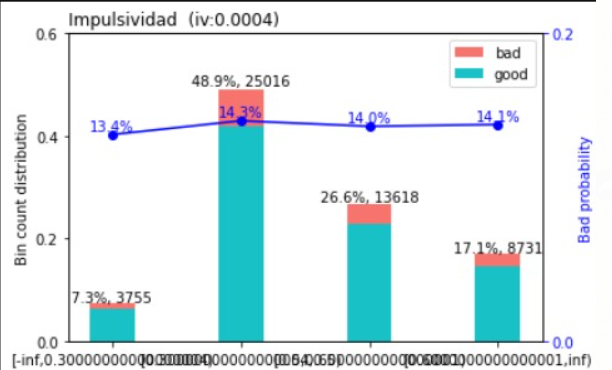
Fuente: Elaboración Propia

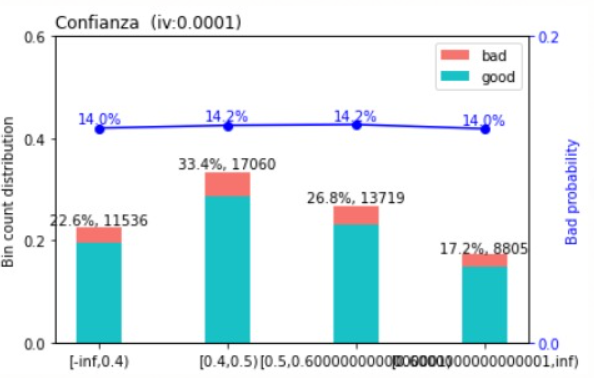
A continuación, se realiza el proceso de feature engineering, seleccionando los atributos de mayor peso y también de mayor Information Value (IV).

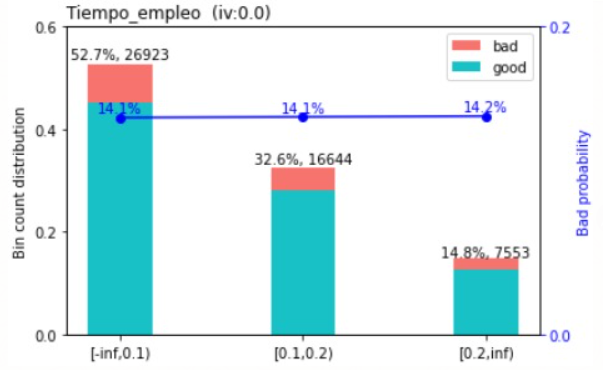












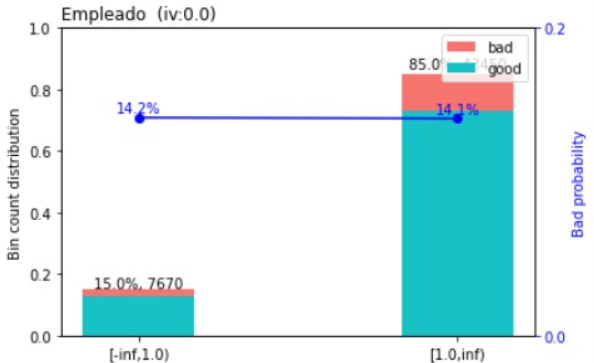


Fig. 7. Gráficos de binarización de variables por WOE

Fuente: Elaboración Propia

De los gráficos mostrados cabe resaltar que todas las variables presentan un “iv” mucho menor a 0.1, el cual es el mínimo para determinar si una variable es predictiva. Además de ello ninguna variable supera el 1% de correlación con la variable target por lo que acorde a los experimentos planteados la estrategia a abordar será encontrar estabilidad sobre los indicadores mas que buscar los mejores indicadores.

# RESULTADOS DE LOS MODELOS APLICADOS

Tras la aplicación de distintas técnicas de balanceo, escalamiento y creación de nuevas variables, se advierte que ningún modelo generado tras los experimentos pudo lograr una mejora simultánea en las métricas usadas, esto debido a la muy baja correlación existente entre las variables originales frente al objetivo (Target).

Por lo antes expresado tuvimos 2 casos recurrentes:

1. Recall mayor a 90%, en este caso el costo de cubrir a este alto porcentaje del incumplimiento real está en línea con predecir también al 90% de casos como incumplimiento, con una efectividad menor o igual a 14%.
2. Efectividad mayor a 90%, en este caso el ser tan certeros tenia como costo el solo atinar a menos del 1% del incumplimiento real.

Ambos casos expuestos son inmanejables en pro de una toma de acción frente al incumplimiento, ya que el primer caso es como notificar a los ejecutivos de la entidad que tomen acción sobre todos los clientes; mientras que en el segundo caso es como notificar que solo se garantiza la detección del 1% de todos los casos de incumplimiento.

Sin embargo, dentro de todas las variantes modeladas también se observa los siguientes puntos:

1. Las variables psicométricas están dentro de las variables significativas para un modelo de Regresión Logística, mientras que no aparecían como importantes para el modelo de Árbol de Decisión
2. Las variantes de regresión Logística con el cambio de punto de corte, de 0.5 a 0.148 escapan de los casos inmanejables descritos líneas arriba.
3. Con el agregado de nuevas variables a partir de trasformaciones polinómicas se logra

Y es así que se opta por elegir como mejor modelo al desarrollado en el experimento “1” con las técnicas de orden “b”, lo que corresponde a un modelo de Regresión Logística con sus variables escaladas en la misma proporción (rango de 0 a 1), entrenado con validación cruzada estratificada de 3 folds con 2 repeticiones de Target, sin balanceo ya que al aplicar esta técnica sobre ajustaba al modelo. En este caso se obtiene la siguiente información:

TABLA VI

RESULTADO SELECCIOANDO

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MÉTRICA | | Dataset de Entrenameinto | | Dataset de validación | |
| recall | | 41.9 % | | 42.5% | |
| efectividad | | 14.4% | | 14.6% | |
|  | |  | |  | |
| accuracy | | 56.6% | | 56.9% | |
| Cobertura de Población a aplicar una acción | | 41.0% | | 40.0% | |

Fuente: Elaboración Propia

Lo resaltante en la tabla anterior es el agregado de “Cobertura de población a aplicar una acción”, lo cual representa todos los clientes predichos como incumplimiento sobre el total de clientes y es añadido para hacer visible que tras el aumento de indicadores en el dataset de validación disminuye la población cubierta (hecho determinante en su elección como mejor modelo).

# CONCLUSIONES

* Las variables psicométricas son significativas en el cálculo del incumplimiento acorde a lo esquematizado en la tabla “V”, donde se les aprecia en el top de variables con mayor peso sobre el modelo seleccionado.
* Con el modelo seleccionado se garantiza capturar el 42% del incumplimiento solo tomando una acción sobre el 40% de los usuarios objetivos, lo cual debe ser sujeto a seguimiento y mejora acorde a la disponibilización de información.
* Tras el cumplimiento de todos los supuestos de la Regresión Logística, este modelo se vuelve el más idóneo para garantizar estabilidad, lo cual, se puede visualizar en la tabla “VI”, donde se aprecian mejores indicadores en el dataset de validación.
* Es necesario la concertación con criterio experto en pro de la búsqueda de nuevas variables que sumen a la predicción del incumplimiento, acorde a las figuras 6 y 7 donde se aprecia que ninguna variable superar los valores mínimos para garantizar buena predicción.

# RECOMENDACIONES

* Se recomienda el aumento de variables psicométricas en pro de contrastar el peso de las mismas sobre los modelos de incumplimiento.
* Se recomienda una re-evaluación sobre los clientes objetivos en pro de la toma de nuevos comportamientos.
* Sumando a los expresado en conclusiones se recomienda el aumento de variables sociodemográficas y de negocio acorde a criterio experto.

References

1. Atiya, A. F. (2001). Bankruptcy Prediction for Credit Risk Using Neural. IEEE TRANSACTIONS ON NEURAL NETWORKS, 929-935.
2. Thomas, L. C., Edelman, D. B., Crook, J. N., (2004), Readings in Credit *Scoring*, Oxford University Press, Oxford.
3. Hsia, D. C. (1978). Credit *Scoring* and the Equal Credit Opportunity act. Hast. Law. J., 30, pp. 371-448.
4. Reichert, A. K., Cho, C. and G. M. Wagner, (1983). An Examination of Conceptual Issues Involved in Developing Credit-*Scoring* Models. J. Bus. Econ. Stat., 1 (2), pp. 101-14.
5. Joanes, D. N. (1993). Reject Inference applied to Logistic Regression for Credit Scoring. IMA J. Math. Appl. Bus. Industry, 5, pp. 35-43.
6. Hand, D. J., and W. E. Henley (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. Journal of the Royal Statistical Society: Series A (Statistics in Society), pp. 160: 523-541.
7. Durand, D. (1941). Risk Elements in Consumer Instalment Financing, Studies in Consumer Instalment Financing. New York: National Bureau of Economic Research.
8. Myers, J. H. and Forgy, E. W. (1963). The Development of Numerical Credit Evaluation Systems. J. Am. Statist. Ass., 58, pp. 799-806.
9. Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 23(4), pp. 589-609.

**[**11] Klinger, B., Khwaja, AI, y del Carpio, C. (2013). Psicometría emprendedora y reducción de la pobreza. Springer Breves en Psicología. Nueva York, NY: Springer New York. Recuperado de <http://link.springer.com/10.1007/978-1-4614-7227-8>.

[12] Colonia, D. (2012). Validez de pronóstico del modelo credit scoring en una entidad microfinanciera, pp. 30.

[13] Gerón, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. O’Reilly Media, Inc.

[14] Damodaran, A. (2007). Probabilistic approaches: scenario analysis decision trees and simulations. Working Paper. Stern School of Business, New York.