Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Ciudad de México

Análisis de métodos multivariados en ciencia de datos

Dr. Andrés Nucamendi

Dr. Iván Ongay

Evidencia del reto

Alan Uriel Merlan Esquivel A01656612

Héctor Hibran Tapia Fernández A01661114

Daniela Jiménez Téllez A01654798

María Fernanda Pérez Ruiz A01742102

Septiembre, 2023.

Resumen

En este documento se trata la problemática respecto al riesgo crediticio en el ámbito de préstamos en México. Se hicieron 3 modelos: análisis factorial, análisis determinante, y regresión logísitica, los cuales predijeron a qué clientes es recomendable hacerles un préstamo o no. De esto se obtuvo que la regresión logística fue la que mejor se acoplaba a la base de datos.

1. Introducción

Actualmente, en México existe un problema respecto al riesgo crediticio en el ámbito de préstamos. De acuerdo con Sebastian Estrada, escritor de la revista *El Economista*, este año se esperan más demandas de préstamos en el país debido a la reciente inflación (Estrada, 2023). Como consecuencia a esto, muchas entidades bancarias necesitan trabajar en sus métodos de selección para ver a qué usuarios se les puede hacer un préstamo y a cuáles no sin causar un daño significativo a sus finanzas. Usualmente, estos mecanismos de selección se basan en características como historial crediticio, adeudos, ingresos, entre otros.

En este proyecto se trabajará con la base de datos "Loan default dataset", la cual contiene una serie de variables que crean un perfil de usuario para así poder proporcionar una posible solución a la problemática anteriormente mencionada. Este conjunto de datos será analizado y limpiado para así poder utilizar diferentes tipos de alogritmos que permiten llegar a una modelo que prediga a qué usuarios se les puede hacer un préstamo y a cuáles no.

Asimismo, para poder obtener resultados se plantean las siguientes preguntas de investigación: ¿qué características tiene un cliente que es mal pagador? y ¿a qué cliente sí conviene hacerle un préstamo?. Igualmente, se tiene la siguiente hipótesis: los no pagadores son aquellas personas con baja solvencia y mayormente hombres.

2. Marco Teórico

Como se mencionó anteriormente, en este proyecto se usaron tres diferentes algoritmos que permitieron llegar a una predicción. A continuación, se da una explicación de qué son y cómo funciona cada uno.

2.1. Análisis factorial

"El análisis factorial es una técnica estadística que se utiliza para explicar un conjunto de variables u observaciones correlacionadas en términos de un número menor de variables no observadas llamadas factores." (Li, 2022.) Este tipo de análisis se utiliza con conjuntos de datos multivariados y su principal propósito es ayudar a disminuir la dimensionalidad de los datos. Esto se logra encontrando patrones entre las variables para así poder agruparlas. La formula del analisis factorial es:

$$Y = \mu + \lambda F + \varepsilon$$

Donde:
Y es la matriz de datos observados (n x p), n represnta el número de observaciones y p el número de variables observadas.
μ represnta al vector de medias de las variables observadas.
λ es la matriz de cargas factoriales (p x m), m es el número de factores a extraer.
F representa la matriz de factores (m x n).
ε represnta la matriz de errores (n x p).

2.2. Análisis discriminante

"El análisis discriminante crea un modelo predictivo para la pertenencia al grupo." (IBM Documentation, 2021.) Su objetivo principal es encontrar una combinación lineal de las variables predictoras que maximice la separación entre los grupos mientras minimiza la variabilidad dentro de cada grupo. La técnica de análisis discriminante propone la determinación de un criterio que permita decidir a qué grupo es que pertenece cierto individuo o variable, a partir de la información que disponemos sobre él cifrada en términos de valores que toman ciertas variables consideradas. No se pretende clasificar a los individuos en grupos, sino que, los grupos ya están previamente construidos y lo que se busca es definir que tiene de específico cada grupo para poder ser capaces de asignar de manera correcta los individuos (variables) a los grupos. La formula del analisis discriminante es:

$$D = XW$$

Donde:

D es el vector de puntuaciones discriminantes.

X representa la matriz de datos (n x p), en donde n es el número de observaciones y p es el número de variables predictoras.

W representa el vector de pesos discriminantes (o coeficientes lineales) (p x 1).

2.3. Regresión logística

La regresión logística se utiliza para observar la relación entre una variable dependiente categórica, la cual puede ser binaria o binomial (como 0 y 1), y una o más variables independientes, a las cuales son les conoce como variables predictorias. Esta usa una función logísitica (igualmente llamada función sigmoide) que permite cambiar la combinación lineal de las variables independientes en una probabilidad. Se ve de la siguiente manera:

$$\frac{1}{1 + e^{-x}}$$

3. Metodología

3.1. Limpieza y preparación de datos

La preparación y limpieza de datos fue vital en el contexto de este reto ya que la calidad de los resultados entregados dependen directamente de la base de datos ya que se busca generar modelos y predicciones confiables y precisos. Para hacer esto, se agregaron variables que fueran caracterísiticas del usuario y se quitaron las que fueran causales del "Status", que es la variable a predecir, para que no fueran a interferir con el análisis. Igualmente, si estas estaban correlacionadas, se tomaba solo una variable (en el caso categórico). Asimismo, debido a que el dataframe tenía muchos valores faltantes (NaN), se contó la cantidad de estos para después se dividirla por status (0 y 1). Si más del 90 % de los datos faltantes pertenecía al status 0 o 1 entonces se rellenaba con un valor que hiciera permitiera que el status original siguiera igual.

3.2. Selección de variables

Una vez teniendo la base de datos limpia, se seleccionaron las variables que son relevantes y que influyen en la decisión crediticia que se busca hacer. Es por esto que se trata de identificar las variables incoherentes o irrelevantes para así poder centrar el análisis en las que tienen más peso, ya que estas ayudan a determinar si el usuario es propenso o no a incumplir con los pagos de su préstamo. En este caso, se ocuparon las variables más comunes como LTB, credit score, credit worthtiness, entre otras, y de igual manera, se agregaron variables que describieran al pagador, como age y gender. De estas variables se descartartaron las que tuvieran una alta correlación entre sí, y las restantes se consideraron no relevantes para el objetivo, por lo que no se ocuparon.

3.3. Análisis discriminante

Para hacer el análisis discriminante se ocupó la librería Sklean en Python. De igual manera, para poder ocupar el modelo se utilizaron las mismas columnas mencionadas anteriormente y "Statusçomo y. Asimismo,

Paso 1. Se crea una instancia del modelo de LDA. Matemáticamente, el LDA busca encontrar las combinaciones lineales de las características que mejor separan dos o más clases de objetos o eventos. Lo hace maximizando la distancia entre las medias de las clases y minimizando la variación dentro de cada clase.

Paso 2. En esta sección, el modelo LDA está siendo entrenado usando un conjunto de datos de entrenamiento (Xtrain para las características y ytrain para las etiquetas). Matemáticamente, el modelo aprende a diferenciar las clases basándose en las relaciones estadísticas entre las características y las etiquetas de las clases.

Paso 3. El modelo ya entrenado está siendo usado para hacer predicciones sobre un conjunto de datos de prueba (Xtest). Matemáticamente, para cada punto en el conjunto de datos de prueba, el modelo calcula la probabilidad de que pertenezca a cada clase y luego asigna la clase con la probabilidad más alta.

Paso 4. Finalmente, se evalua el rendimiento del modelo usando una matriz de confusión y un informe de clasificación, que se calculan comparando las etiquetas verdaderas (y.test) con las etiquetas predichas (ypred).

3.4. Análisis factorial

Para hacer el análisis factorial se necesitó separar las variables que se escogieron en las numéricas. Se ocupó una rotación oblicua debido a las características de los datos, en este caso "pro min". De los resultados se hizo un cálculo de dos factores para reducir la dimensionalidad de las variables, lo que dio paso a hacer una regresión logísitica.

3.5. Regresión logística

En el caso de la regresión logísitica, se hizo uso de los factores que se sacaron en el análisis factorial y las variables categóricas. Con estos, se hizo un análisis de máximo 1000 iteraciones, lo que dio como resultado un f-score alto.

4. Resultados y análisis

4.1. Resultados con el análisis discriminante

Del análisis discriminante se obtuvo la siguiente matriz de confusión:

Igualmente, se obtuvieron los siguientes resultados:

4.2. Resultados con el análisis factorial

Del análisis factorial se decidió hacer 2 factores para reducir la dimensionalidad de las variables. A continuación se muestran los resultados:

Donde el factor 0 es: "Factor de solvencia" contiene las variables de 'loan_amount' e 'income', y engloba la relación entre el monto del préstamo y los ingresos, lo cual indica la capacidad de cierta persona para cumplir con su compromisos financieros.

Y el factor 1 es: "Factor de Costo de Financiamiento Inicial" resalta la relación que hay entre los cargos iniciales y la tasa de interés, indicando el costo inicial de financiamiento que la persona debe pagar.

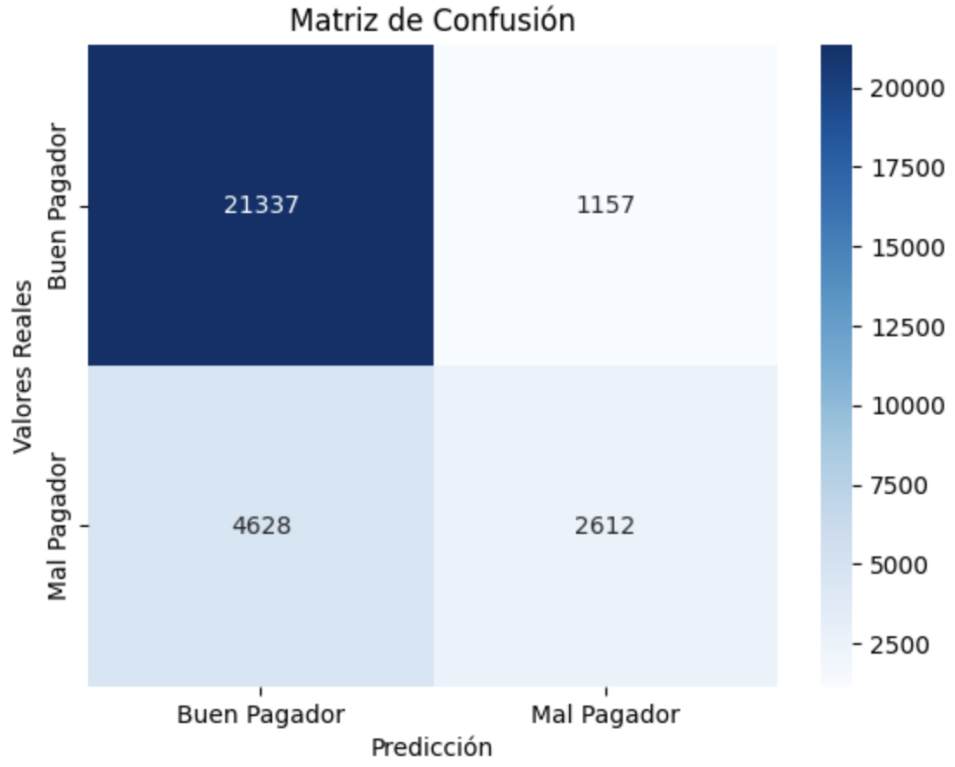

Figura 1: Matriz de confusión del análisis discriminante

4.3. Resultados con la regresión logísitica

Los resultados que se obtuvieron con la regresión logísitica fueron los mejores. De lo anterior se obtuvo la siguiente matriz de confusión:

Igualmente, se obtuvo un **F-score: 0.94387**. Y los coeficientes de la regresión:

De esto se pudieron hacer las siguientes gráficas para describir a un mal pagador:

Y la siguiente gráfica describe a un buen pagador:

	precision	recall	f1-score	support
0	0.82	0.95	0.88	22494
1	0.69	0.36	0.47	7240
accuracy			0.81	29734
macro avg	0.76	0.65	0.68	29734
weighted avg	0.79	0.81	0.78	29734

Figura 2: Resultados del análisis discriminante

	0	1	Comunalidad
loan_amount	0.453521	-0.020348	2.056817e-01
Upfront_charges	-0.002152	0.432832	4.629657e-06
LTV	-0.068762	-0.087726	4.728207e-03
rate_of_interest	0.000264	0.703806	6.981835e-08
dtir1	-0.270822	-0.067505	7.334440e-02
Credit_Score	0.003243	-0.008033	1.051942e-05
income	0.997451	0.000573	9.949092e-01
Var	1.278679	0.695418	1.974097e+00
%Var	0.106557	0.057951	1.645080e-01

Figura 3: Cargas factoriales del análisis factorial

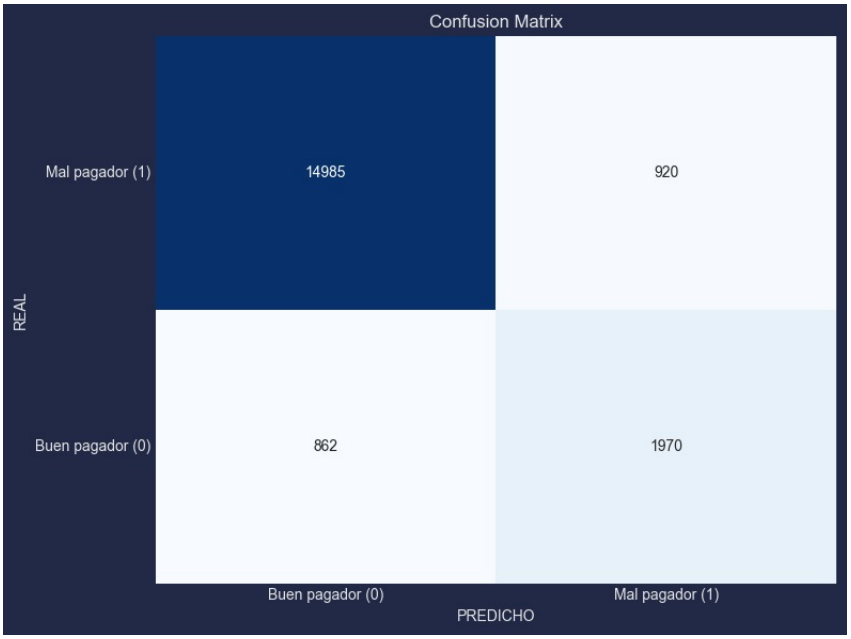

Figura 4: Matriz de confusión de la regresión logística

	0	1	Credit_Worthiness	loan_amount	income	Credit_Score	LTV	Gender	loan_purpose	open_credit	Upfront_charges	age	dtir1	intercept
0	-0.000054	0.001982	0.000227	-0.000003	-0.000027	0.000712	0.005935	-0.000468	-0.002181	0.000022	-3.33229	0.001627	0.011103	0.000122

Figura 5: Coeficientes de la regresión logísitica

Figura 6: Gráfica de barras que describe a un mal pagador

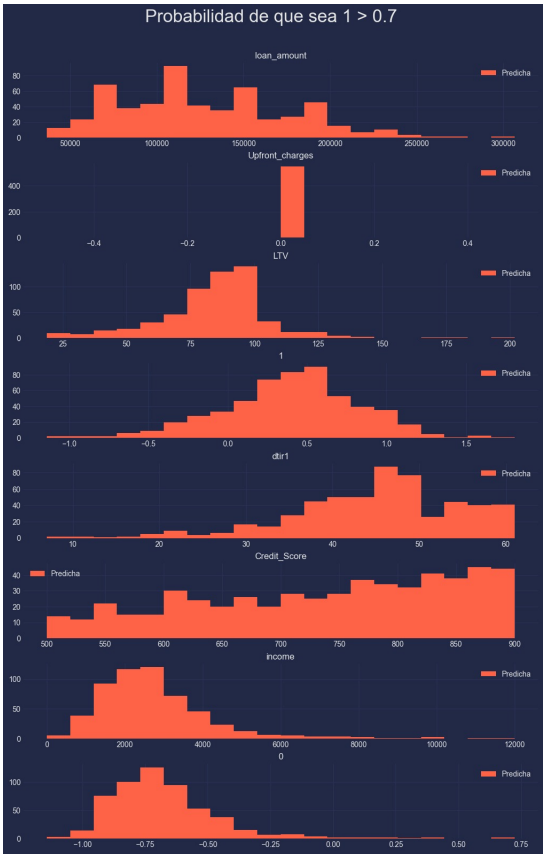

Figura 7: Histogramas que describen a un mal pagador

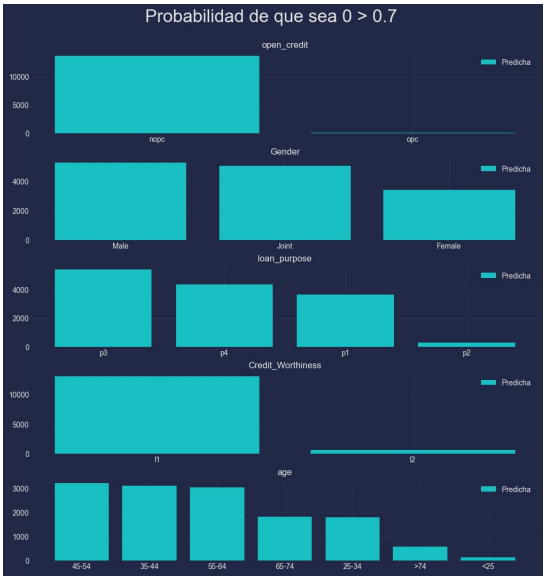

Figura 8: Gráfica de barras que describen a un buen pagador

5. Discusión

A continuación se explorarán trabajos similares y estudios anteriores que se relacionancon la predicción de riesgo crediticio, y se comprobará si los resultados obtenidos se alinean con las tendencias identificadas por otros autores o si se destacan por alguna razón particular. Usualmente "las entidades bancarias utilizan el coeficiente o ratio de endeudamiento para determinar si un particular o una empresa puede permitirse pagar las cuotas mensuales. Dicho porcentaje oscila entre el 35 y 40 porciento de los ingresos totales y supone un límite más allá del cual no sería seguro para un banco prestar una cantidad determinada de dinero a un cliente."(BBVA, 2015) El ing. Paresh Khandelwal describe la idea de que el riesgo de los prestamos crediticios puede evaluarse con modelos de opciones.

En Ecuador, Jaime Pérez destaca que el acceso al crédito financiero es una herramienta crucial para el desarrollo de proyectos productivos en su país. Sin embargo, se enfrenta a un alto nivel de incumpliento en los pagos de los préstamos, es por eso que usa el modelo de Regresión Logística Binaria para estimar la probabilidad de morosidad crediticia, logrando con dicho modelo una predicción del 70 por ciento proporcionando confiabilidad para tomar decisiones sobre la aprobación de créditos.

6. Conclusiones

Finalmente, se puede concluir que el modelo que mejor se acopló a la base de datos fue el de regresión logística. Para poder lograr este, se tuvo que hacer un análisis factorial para poder reducir la dimensionalidad de las variables y que fuera más óptimo el resultado. De esta regresión se pudo inferir que las características de un mal pagador son: nopc,hombre,p3 (compra de un vehiculo),I1 (indica un nivel de solvencia crediticia superior o excelente) y de edad entre 45 y 54 años. Por otro lado, las caracterísiticas de un buen pagador son: nopc, hombres,p3 (compra de un vehiculo), I1 (indica un nivel de solvencia crediticia superior o excelente) y de edad entre 45 y 54 años. Tiene logica que tengamos las mismas variables como buen y mal pagador ya que las compara para encasillar a la persona. De igual manera, se puede decir que la hipótesis se acepta, sin embargo hubieron más variables que pudieron corroborar el perfil de un mal pagador. A pesar de esto, se cree que se obtuvieron buenos resultados, y que tal vez con más tiempo y más detalles de la base de datos, se podría tener

un modelo más detallado.

7. Referencias

Li, Haksun. (2022). Numerical Methods Using Java - For Data Science, Analysis, and Engineering - 12.8 Factor Analysis. (pp. 843). Apress, an imprint of Springer Nature. Retrieved from https://app.knovel.com/hotlink/pdf/id:kt01349T51/numerical-methods-using/factor-analysis

Fortino, Andres. (2023). Data Mining and Predictive Analytics for Business Decisions - A Case Study Approach - 7.14 Exercise 7.3 - PassClass Case Study. Mercury Learning and Information. Retrieved from https://app.knovel.com/hotlink/pdf/id:kt0138V3S1/data-mining-predictive/exercise-7-3-passclass

BBVA. (2015, April 29). ¿Cómo aprueban o deniegan las entidades financieras las solicitudes de préstamo? BBVA NOTICIAS. https://www.bbva.com/es/como-aprueban-o-deniegan-las-entidades-financieras-las-solicitudes-de-prestamo-y-ii/

Intuition of Mathematical Modelling of Loans as Options. (2021). https://pulse/intuition-mathematical-modelling-loans-options-paresh-khandelwal

Pérez, J. (n.d.). LA REGRESIÓN LOGÍSTICA COMO MODELO DE PREDICCIÓN DEL RIESGO CREDITICIO EN LAS ORGANIZACIONES DE LA ECONOMÍA SOCIAL Y SOLIDARIA (The logistic regression as a model of prediction of credit risk in organizations of the social and solidarity economy). https://www.uv.mx/iiesca/files/2018/03/23CA201702.pdf

I.M.,Lejarza,2018 analisis discriminate

https://www.uv.es/mlejarza/actuariales/tam/discriminante.pdf

Análisis Discriminante Santiago de la Fuente Fernández. (n.d.). https://www.estadistica.net/Master-Econometria/AnalisisDiscriminante.pdf