Credit scoring

Índice

1. Introducción

Los modelos de credit scoring se han convertido en una norma en la banca moderna, dado el incremento en el número de solicitudes de crédito, una regulación en la industria financiera más estricta y el desarrollo de la tecnología, han conducido al desarrollo de modelos estadísticos que ayudan en el proceso de otorgamiento de crédito.

El objetivo de este trabajo es investigar que es el credit scoring, como funciona, como se lleva a cabo la creación de un modelo de credit scoring y la implementación de un modelo de credit scoring en un conjunto de datos reales usando el lenguaje de programación Python. El trabajo se divide en 5 secciones.

En la *sección 1, Introducción* se explica que es el riesgo de crédito, que es el credit scoring y cuáles son los principales pasos para la creación de un modelo. En la *sección 2, Metodología* se describe el modelo estadístico más utilizado para la creación de modelos de credit scoring, la regresión logística, también se presentan los métodos usados para la selección de variables y el método usado para transformar de los resultados de la regresión en un formato de scorecard, así como las métricas usadas para evaluar el ajuste y desempeño del modelo. En la *sección 3, Descripción de datos* se presentan los datos que usamos para la creación de un modelo de credit scoring. En la *sección 4, Resultados* se analiza el desempeño del modelo creado utilizando los datos mencionados anteriormente, así como un análisis de por que obtuvimos estos resultados. Finalmente en la sección *sección 5, Segmentación* se investiga el proceso de segmentación en los modelos de credit scoring, así como una comparación entre el desempeño del modelo usando segmentación.

1.1. Credit scoring

Antes de definir que es el credit scoring es necesario primero definir y entender que es un crédito. El crédito es una operación financiera en la que un acreedor (generalmente una institución financiera) presta dinero a un deudor, el cual se compromete a pagar el monto prestado más un cierto interés en un periodo determinado. El riesgo de crédito se presenta cuando el deudor esta poco dispuesto o imposibilitado de cumplir con su obligación de pago.Su efecto se mide por el costo de la reposición de flujos de efectivo si el deudor incumple.

La calificación crediticia o credit scoring es el nombre usado para describir el proceso y los modelos estadísticos usados para determinar qué tan probable que un solicitante de crédito incumpla con su promesa de pago. Su objetivo es crear una sola medida de riesgo a partir de un conjunto de factores de riesgo que nos permita clasificar a los solicitantes de crédito como buen o mal riesgo. Los métodos estadísticos con los que estimamos estas probabilidades de incumplimiento son conocidos como clasificadores. Estos modelos estadísticos transforman datos

relevantes de los solicitantes de crédito en una medida numérica que ayuda a decidir si se debe otorga el crédito. La decisión de aceptar o rechazar otorgar el crédito se toma comparando la probabilidad de incumplimiento estimada con un umbral que se considere adecuado.

Un scorecard consiste en un grupo de características, determinadas estadísticamente, que son predictivas en poder separar clientes buenos y malos. Las características seleccionadas provienen de datos disponibles al momento de la solicitud del crédito. A cada atributo se le asigna un puntaje basado en análisis estadístico. El puntaje total para un aplicante es la suma de los puntos individuales para cada atributo. La mayoría de los scorecard asumen tener una relación monótona entre el puntaje y la probabilidad de ser buen cliente. Los clientes buenos tienden a tener puntajes mas altos y son considerados menos riesgosos, puntajes bajos corresponden a clientes que tienen una mayor probabilidad de ser malos.

Algunos de los métodos de clasificación conducen a un scorecard, donde a cada característica se le es dada un puntaje, el puntaje total determina si el cliente es bueno o malo. Otros métodos no conducen a scorecards, mas bien estiman directamente la probabilidad de ser bueno o no.

1.2. Desarrollo de un modelo de credit scoring

El proceso de desarrollo e implementación de un modelo de credit scoring se divide en varias etapas, las cuales se describen brevemente a continuación.

- Entender el plan de negocio: Definir el objetivo del modelo y para que va a ser utilizado.
- Definir la variable objetivo: Definir el evento de interés, usualmente el evento de interés es definido como “malo” o “default”, el default es ocurre cuando la institución financiera considera que es poco probable que el prestatario cumpla con su obligación o cuando el prestatario tiene mas de 90 días de atraso.
- Datos: Esta etapa consta de la recolección, almacenamiento, limpieza y transformación de los datos.
- Ajustar y optimizar el modelo: Seleccionar el modelo estadístico y evaluar su poder predictivo.
- Generalización: Probar la habilidad predictiva del modelo en una muestra no utilizada durante la fase de entrenamiento.
- Monitoreo: Una vez que el modelo ha sido desarrollado e implementado, es necesario revisar que el modelo siga prediciendo correctamente.

2. Metodología

Un supuesto clave para construir un modelo de credit scoring es que el futuro se parece al pasado. Mediante el análisis del comportamiento previo de los clientes a los que se les otorgo el crédito y ya conocemos si fueron buenos o malos, es posible aprender y predecir como se comportaran los clientes futuros. El comportamiento de los clientes, que denotaremos G para bueno y B para malo, es la variable objetivo que nos interesa analizar, así como su relación con las características de los solicitantes.

Sea $\mathcal{X} = (x_1, x_2, ..., x_m)$ un conjunto de datos con m instancias (o muestras), done cada instancia representa un crédito que fue otorgado en el pasado y es descrito por d características $x_i = (x_{i1}, x_{i2}, ..., x_{id})$. Si sabemos que el comportamiento de la instancia i fue y_i donde $y_i \in \{G, B\}$. Entonces el conjunto $\mathcal{D} = \{(x_i, y_i)|x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, 2, ..., m\}$ llamado conjunto de entrenamiento, es la materia prima necesaria para construir un modelo de credit scoring. Un modelo credit scoring estima las probabilidades de que un solicitante de crédito con un vector de características $\mathbf{x} = (x_1, x_2, ..., x_d)$ se comporte relativamente bien durante el préstamo o se comporte mal (incumpla con su compromiso de pago) denotados por $P(G|\mathbf{x})$ y $P(B|\mathbf{x})$ respectivamente.

Nuestra tarea consiste en crear un modelo que nos permita estimar las probabilidades $P(.|\mathbf{x})$ a partir de un conjunto de datos $\mathcal{D}$ y a partir de estas probabilidades encontrar una regla que nos permita dividir el espacio generado por $\mathcal{X}$ en dos subespacios A_G y A_B, donde los solicitantes cuyo vector de características $\mathbf{x} \in A_G$ tienen una mayor probabilidad de comportarse bien durante el periodo del crédito. Además, buscamos que el modelo generalice, es decir, para un nuevo conjunto de datos $\mathcal{D}' = \{(x_i, y_i)|x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = m+1, m+2, ..., m+m'\}$ llamado conjunto de prueba, el modelo sea capaz de clasificar correctamente estas nuevas observaciones.

2.1. Regresión logística

La regresión logística, al igual que la regresión lineal, es una técnica estadística usada para analizar la relación entre una variable dependiente y una o más variables independientes. La principal diferencia entre la regresión logística y la regresión lineal es que la variable dependiente es una variable binaria $y \in \{0, 1\}$. Por lo que el valor real predicho por una regresión lineal $z = b + \mathbf{w}^T\mathbf{x}$ necesita ser convertido, ya sea, en 0 o 1.

Sean p la probabilidad de que una muestra $\mathbf{x}$ pertenezca a la clase 1 y $q = 1 - p$ la probabilidad de que pertenezca a la clase 0, la razón $\frac{p}{1-p}$ es llamada momio e indica la probabilidad relativa de que la muestra $\mathbf{x}$ pertenezca a la clase 1. Tomando el logaritmo del momio $\log \frac{p}{1-p}$ obtenemos el log-momio (función logit). Si usamos una regresión lineal para aproximar el valor del log-momio de

la clase real.

$$\log \frac{p}{1-p} = b + w_1 x_1 + w_2 x_2 + ... + w_d x_d = b + \mathbf{w}^T \mathbf{x}$$

Resulta que la probabilidad estimada de que una muestra $\mathbf{x}$ pertenezca a la clase 1 esta dada por la función sigmoide.

$$P(Y = 1|\mathbf{x}, b, \mathbf{w}) = p = \frac{1}{1 + e^{-(b+\mathbf{w}^T \mathbf{x})}}$$

La estimación de los parámetros b,w_1,...,w_d es llevada a cabo utilizando el método de máxima verosimilitud. Este método estima los parámetros que maximizan la probabilidad de obtener la muestra observada.

Reescribiendo b y $\mathbf{w}$ como $\mathbf{w}' = (\mathbf{w}, b)$ y representando las variables independientes como una matriz $\mathbf{X}$ de dimensiones $m \times (d+1)$ y la variable dependiente como un vector columna y. La función de verosimilitud a maximizar está dada por:

$$L(\mathbf{w}') = \prod_{i=1}^{m} P(y_i = 1|\mathbf{x_i}, \mathbf{w}')^{y_i} * P(y_i = 0|\mathbf{x_i}, \mathbf{w}')^{1-y_i}$$

Si tomamos el logaritmo de la función de verosimilitud y obtenemos la función de log-verosimilitud $l(\mathbf{w}')$ la cual es más fácil maximizar.

$$l(\mathbf{w}') = \sum_{i=1}^{m} (y_i * \mathbf{w'}^T \mathbf{x} - \log(1 + e^{\mathbf{w'}^T \mathbf{x}}))$$

La solución al problema de maximizar $l(\mathbf{w}')$ es la misma que la del problema de minimizar $-l(\mathbf{w}')$, por lo que los valores óptimos $\mathbf{w}'$ puede ser encontrados aplicando técnicas diferentes métodos de optimización tales como el método de Newton o el método de gradiente descendiente.

2.1.1. Regularización

La regularización es un método bastante útil para manejar la colinealidad, filtrar el ruido de los datos y ayuda a prevenir el sobreajuste, los métodos de regularización mas utilizados son la regularización L_1 y la regularización L_2.

La regularización L_2, comúnmente conocido como regularización Ridge contrae los coeficientes del modelo mediante una penalización en su tamaño. Se basa en incrementa la función de costo por un factor igual a la suma de los cuadrados de los coeficientes multiplicados por un parámetro de regularización λ el cual controla el monto de contracción, a mayores valores de λ mayor será la contracción.

$$min_{\mathbf{w}'}(-l(\mathbf{w}') - \lambda \times \sum_{i=1}^{d} w_i^2)$$

La regularización L_1, comúnmente conocido como regularización Lasso contrae los coeficientes del modelo mediante una penalización en su tamaño. Se basa en incrementa la función de costo por un factor igual a la suma de los valores absolutos de los coeficientes multiplicados por un parámetro de regularización λ el cual controla el monto de contracción, a mayores valores de λ mayor será la contracción. Al utilizar valores grandes de λ algunos de los coeficientes pueden ser cero, por lo que también es usada como una técnica de selección de variables.

$$min_{\mathbf{w}'}(-l(\mathbf{w}') - \lambda \times \sum_{i=1}^{d} |w_i|)$$

La principal diferencia entre la regularización L_1 y L_2 es que los coeficientes obtenidos por la regularización L_2 están cerca del cero, mientras que los coeficientes obtenidos por la regularización L_1 pueden ser cero.

2.2. Selección de variables

El análisis de variables investiga la relación entre las variables independientes, que uno quiere probar su capacidad predictiva y la variable dependiente que se espera predecir. La reducción de variables permite reducir las variables usadas para construir el modelo con la menor perdida de información además de mantener las variables con un mayor predictivo.

El objetivo es seleccionar las variables con las que se puede crear un mejor modelo, reducir el número de variables puede resultar en un modelo más estable y que puede generalizar mejor.

2.2.1. Weight of evidence

Los Weight of evidence (WoE) mide la fortaleza de un atributo en la separación de clientes buenos y malos convirtiendo el riesgo asociado en una escala lineal que es fácil de interpretar por los humanos.

Para una variable categórica con r atributos, sean g_i y b_i el número de buenos y malos en el atributo i, el WoE del atributo i es:

$$WoE_i = \log((\frac{g_i}{\sum_{i=1}^{r} g_i})/(\frac{b_i}{\sum_{i=1}^{r} b_i}))$$

Un WOE negativo implican que un atributo particular está aislando una mayor proporción de malos que buenos.

2.2.2. Information value

El information value (IV) tiene sus orígenes en la teoría de la información (divergencia de Kullback-Leibler entre las distribuciones de buenos y malos) y es calculado:

$$IV = \sum_{i=1}^{r} (\frac{g_i}{\sum_{i=1}^{r} g_i} - \frac{b_i}{\sum_{i=1}^{r} b_i}) * WoE_i$$

Es una medida del poder discriminatorio de la variable, a mayor IV, los atributos de la variable distinguen mejor entre los buenos y malos. Una variable con un IV mayor o igual a 0.1 es considerada informativa y adecuada para usarse en el modelo.

En general se pueden considerar los valores del IV como:

- Sin poder predictivo: $IV < 0{,}02$
- Débilmente predictivo: $0{,}02 \leq IV < 0{,}1$
- Moderadamente predictivo: $0{,}1 \leq IV < 0{,}3$
- Fuertemente predictivo: $0{,}3 \leq IV$

2.3. Regresión logística en credit scoring

Dada su simplicidad y buen desempeño, la regresión logística es uno de los modelos de clasificación mas utilizados para credit scoring. Además, este modelo pude ser transformado fácilmente en un formato interpretable basado en puntos de crédito.

Suponga que tenemos un modelo logístico donde todas las variables independientes han sido codificadas utilizando la transformación WoE.

$$P(Y = 1|b, \mathbf{w}) = \frac{1}{1 + e^{-(b + w_1 WoE_1 + w_2 WoE_2 + ... + w_d WoE_d)}}$$

Este modelo puede ser expresado de manera lineal a través del log-momio.

$$\log(\frac{P(Y = 1|b, \mathbf{w})}{P(Y = 0|b, \mathbf{w})}) = b + w_1 WoE_1 + w_2 WoE_2 + ... + w_d WoE_d$$

La relación ente los puntajes y los log-momios se puede expresar de una manera lineal.

$$score = offset + factor * ln(odds)$$

Utilizando un momio (*odds*) a un cierto puntaje (*score*) y unos puntos para duplicar los momios (points to double the odds, *pdo*), los valores de *factor* y *offset* pueden ser resolviendo el sistema de ecuaciones.

$$score = offset + factor * ln(odds)$$

$$score + pdo = offset + factor * ln(2 * odds)$$

Resolviendo para *pdo*.

$$pdo = factor * ln(2)$$

Y finalmente obtenemos los valores de *factor* y *offset* como:

$$factor = \frac{pdo}{\log(2)}$$

$$offset = score - factor * ln(odds)$$

Una vez que estos valores son calculados el puntaje de crédito se calcula como:

$$\begin{aligned} score &= offset + factor * ln(odds) \\ &= offset + factor * (\sum_{i=1}^{N}(WoE_i * w_i) + b) \\ &= offset + factor * (\sum_{i=1}^{N}(WoE_i * w_i + \frac{b}{N})) \\ &= \sum_{i=1}^{N}(\frac{offset}{N} + factor(WoE_i * w_i + \frac{b}{N})) \end{aligned}$$

El puntaje final es comparado con un punto de corte que ayuda a predecir si es mas probable que el cliente sea bueno o malo. La principal ventaja de este puntaje de crédito es su interpretabilidad, podemos ver cuales son las categorías más riesgosas y como contribuyen al puntaje de crédito.

2.4. Medidas de desempeño

La proporción de muestras clasificadas incorrectamente con respecto al número total de muestras es llamada la tasa de error (error rate), uno menos la tasa de error es llamada exactitud (accuracy). El error calculado en el conjunto de entrenamiento es llamado error de entrenamiento o error empírico, el error calculado en nuevos datos es llamado error de generalización. Nuestro objetivo es crear modelos que tangan un error de generalización pequeño, sin embargo, en la práctica solo es posible minimizar el error de entrenamiento. Si un modelo tiene un error de entrenamiento bajo y un error de generalización alto, decimos que el modelo esta sobreajustado (overfitting), en el caso contrario, si un modelo

tiene un error de entrenamiento alto, decimos que el modelo esta subajustado (underfitting).

Podemos evaluar el error de generalización a través de un experimento de prueba, usando un conjunto de prueba, podemos estimar la habilidad de clasificación en nuevos datos y usar este error de prueba como una aproximación al error de generalización. Lo mas usual es dividir un conjunto de datos $\mathcal{D} = \{(x_i, y_i)|x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, 2, ..., m\}$ en dos subconjuntos disjuntos: el conjunto de entrenamiento S y el conjunto de prueba T, tales que $D = S \cup T$ y $S \cap T = \varnothing$. Usamos el conjunto de entrenamiento S para entrenar el modelo y para calcular el error de entrenamiento y utilizamos el conjunto de prueba T para estimar el error de generalización.

2.4.1. Matriz de confusión

La tasa de error y la exactitud son frecuentemente usadas, pero no son adecuadas para todas las tareas de clasificación, por ejemplo, si deseamos conocer el número de clientes buenos que fueron clasificados como malos y viceversa, es necesario definir otra medida de desempeño que nos ayude a contestar este tipo de preguntas.

En problemas de clasificación binaria, hay cuatro combinaciones entre los valores reales y los valores predichos por el modelo, llamados verdadero positivo TP, falso positivo FP, verdadero negativo TN y falso negativo FN, con $TP + FP + TN + FN = m$ (m es el número de muestras en el conjunto de datos), que pueden ser representados mediante la matriz de confusión.

	Predicción negativa	Predicción positiva	Total
Real negativa	TN	FP	N
Real positiva	FN	TP	P
Total	N'	P'	m

- Verdaderos positivos TP: El número de muestras positivas que fueron correctamente clasificadas.
- Verdadero negativo TN: El número de muestras negativas que fueron correctamente clasificadas.
- Falso positivo FP: El número de muestras negativas que fueron clasificadas como positivas.
- Falso negativo FN: El número de muestras positivas que fueron clasificadas como negativas.

El número de clasificaciones correctas es $TP + TN$, el número de clasificaciones incorrectas es $FP + FN$.Con esta notación podemos calcular la tasa de error y la exactitud como $\frac{FP+FN}{m}$ y $\frac{TP+TN}{m}$.

Otras medidas de desempeño que se pueden deducir de la matriz de confusión son la sensibilidad y la precisión. La sensibilidad o recuperación (sensitivity o recall) también es conocida como tasa de verdaderos positivos es una medida de completitud que intenta responder a la pregunta ¿Qué proporción de positivos reales se clasificaron correctamente? y está definida como $R = \frac{TP}{TP+FN} = \frac{TP}{P}$. La precisión (precision) es una medida de exactitud que intenta responder a la pregunta ¿Qué proporción de clasificaciones positivas fue correcta? y está definida como $P = \frac{TP}{TP+FP} = \frac{TP}{P'}$.

2.4.2. ROC y AUC

En ocasiones las predicciones de los modelos de clasificación están dadas en forma de probabilidades o valores reales, las predicciones de las clases se realizan comparando estos valores con un cierto umbral. La curva *ROC* (Receiver Operating Characteristics) es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación usando la tasa de falsos positivos FPR en el eje x y la tasa de verdaderos positivos TPR en el eje y.

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{N}$$

Cada valor de umbral produce un punto (FPR, TPR) si variamos el valor umbral (de $(-\infty, \infty)$ en el caso de valores reales o de $[0, 1]$ en caso de probabilidades) construimos la curva ROC, sin embargo, es imposible evaluar (FPR, TPR) para todos los valores posibles de umbral. Por lo que es necesario una técnica que nos permita estimar la curva ROC de una manera más eficiente.

Sean m^+ y m^- el número de muestras positivas y negativas respectivamente, si ordenamos los valores predichos por el modelo de acuerdo a su valor y establecemos el umbral lo más grande posible (predecimos todas las muestras como negativas) obtenemos el punto $(0, 0)$. Sea (x, y) el punto previo de la curva, si la muestra actual es verdadera positiva establecemos el punto actual de la curva como $(x, y + \frac{1}{m^+})$, por otro lado si la muestra actual es falsa positiva establecemos el punto actual de la curva como $(x + \frac{1}{m^-}, y)$. Finalmente si establecemos el umbral lo más pequeño posible (predecimos todas las muestras como verdaderas) obtenemos el punto $(1, 1)$. Uniendo los puntos anteriores obtenemos la curva ROC.

La curva ROC es una representación del desempeño del modelo. Una manera de comparar modelos es calcular el área bajo la curva ROC, esta área es llamada AUC (Area Under ROC Curve), el AUC puede ser estimado mediante:

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i)(y_{i+1} + y_i)$$

El AUC representa la habilidad de un modelo para separar las clases positivas de las clases negativas, cuanto mayor sea su valor, mayor será su habilidad de separación. También podemos entender el AUC como la probabilidad de que un clasificador esté más seguro de que un ejemplo positivo elegido al azar sea realmente positivo que de que un ejemplo negativo elegido al azar sea positivo.

2.4.3. KS

La prueba de Kolmogorov–Smirnov KS es una prueba de bondad de ajuste usada para comparar si una distribución observada desconocida sigue una distribución teórica conocida. El estadístico de la prueba es calculada tomando la máxima distancia vertical entre las curvas (las funciones de distribución acumuladas):

$$KS = max_x|F_n(x) - F(x)|$$

El estadístico KS toma valores entre 0 y 1, un valor de 0 indica que las distribuciones son idénticas (el modelo no distingue entre clientes buenos y malos), mientras que un valor de 1 indica que las curvas son muy diferentes (el modelo les asigna un puntaje menor a los clientes malos que a los clientes buenos).

3. Descripción de datos

En esta sección hablaremos sobre los datos usados *Give Me Some Credit*, los cuales fueron usados para crear un modelo de credit scoring. Give Me Some Credit es una competencia creada por *Kaggle*, la cual busca que los participantes mejoren el "estado del arte" de la calificación crediticia, mediante la predicción de la probabilidad de que alguien experimente dificultades financieras en los próximos dos años.

El objetivo de esta competencia es construir un modelo que los prestatarios puedan usar para ayudarlos a tomar las mejores decisiones financieras.

Los datos consisten en 150,000 préstamos otorgados que son descritos por 11 variables. Los datos fueron obtenidos de Kaggle y se encuentran disponibles en la página: https://www.kaggle.com/competitions/GiveMeSomeCredit/data

Las variables son descritas en la siguiente tabla.

Variable	**Descripción**
SeriousDlqin2yrs	La persona experimentó 90 días de morosidad o más
RevolvingUtilizationOfUnsecuredLines	Saldo total en tarjetas de crédito y líneas de crédito personales, excepto bienes raíces y sin deuda a plazos, como préstamos para automóviles, dividido por la suma de los límites de crédito
age	Edad del prestatario
NumberOfTime30-59DaysPastDueNotWorse	Número de veces que el prestatario ha estado atrasado entre 30 y 59 días, pero no peor en los últimos 2 años
DebtRatio	Pagos mensuales de deuda, pensión alimenticia, costos de vida divididos por el ingreso bruto mensual
MonthlyIncome	Ingreso mensual
NumberOfOpenCreditLinesAndLoans	Número de préstamos abiertos (a plazos, como préstamos para automóviles o hipotecas) y líneas de crédito (por ejemplo, tarjetas de crédito)
NumberOfTimes90DaysLate	Número de veces que el prestatario ha estado en mora por 90 días o más
NumberRealEstateLoansOrLines	Número de préstamos hipotecarios y de bienes raíces, incluidas las líneas de crédito con garantía hipotecaria
NumberOfTime60-89DaysPastDueNotWorse	Número de veces que el prestatario ha estado atrasado entre 60 y 89 días, pero no peor en los últimos 2 años
NumberOfDependents	Número de dependientes en la familia que se excluyen a sí mismos (cónyuge, hijos, etc.)

3.1. Variable objetivo

El conjunto de datos contiene la variable **SeriousDlqin2yrs** la cual representa si la persona que recibió el préstamo experimentó 90 días de morosidad o más. Representando los clientes que experimentaron mora como 1 (malos) y los demás como 0 (buenos).Esta es la variable objetivo que estamos intentando predecir.

De los 150,000 datos 139,974 (93.31 %) son buenos clientes, mientras que los 10,026 (6.69 %) restantes son malos.

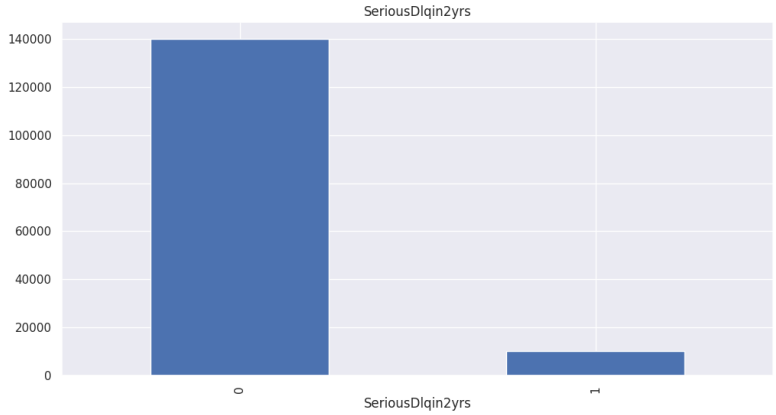

Figura 1: Variable SeriousDlqin2yrs

3.2. Variables predictoras

Las 10 variables restantes son utilizadas como variables predictoras (o explicativas) candidatas para ser usadas en el modelo. Dada la naturaleza de estas variables y su cardinalidad, las variables fueron clasificadas como variables discretas o continuas.

Variable	**Tipo**	**Cardinalidad**
RevolvingUtilizationOfUnsecuredLines	Continua	125,728
DebtRatio	Continua	114,194
MonthlyIncome	Continua	13,594
age	Discreta	86
NumberOfOpenCreditLinesAndLoans	Discreta	58
NumberRealEstateLoansOrLines	Discreta	28
NumberOfTimes90DaysLate	Discreta	19
NumberOfTime30-59DaysPastDueNotWorse	Discreta	16
NumberOfTime60-89DaysPastDueNotWorse	Discreta	13
NumberOfDependents	Discreta	13

3.2.1. Variables continuas

Las variables continuas se caracterizan por su alta cardinalidad y por la presencia de valores extremos. Las variables presentan una asimetría positiva, lo que indica que la mayoría de los valores extremos se encuentran a la derecha de la distribución (cerca del valor máximo).

Estadística	RevolvingUtilizationOfUnsecuredLines	DebtRatio	MonthlyIncome
count	150,000	150,000	120,269
mean	6.05	353.01	6,670.22
std	249.76	2,037.82	14,384.67
min	0.0	0.0	0.0
25 %	0.03	0.18	3,400
50 %	0.15	0.37	5,400
75 %	0.56	0.87	8,249
max	50,708	329,664	3,008,750
missing	0	0	29,731

Podemos detectar posibles valores extremos usando los z-scores (un z-scores es considerado extremo si $|z| > 3$), usando esta medida obtenemos que 1,169 de las observaciones contienen extremos, esto se puede observar en los histogramas y las gráficas de dispersión de las variables, así como la baja correlación entre ellas.

3.2.2. Variables discretas

Las variables discretas se caracterizan por su baja cardinalidad y porque unos pocos valores representan gran parte de la densidad de la variable.

	count	unique	Moda	Frecuencia
NumberOfTime30-59DaysPastDueNotWorse	150000.0	16.0	0.0	126018.0
NumberOfTimes90DaysLate	150000.0	19.0	0.0	141662.0
NumberRealEstateLoansOrLines	150000.0	28.0	0.0	56188.0
NumberOfTime60-89DaysPastDueNotWorse	150000.0	13.0	0.0	142396.0
NumberOfDependents	146076.0	13.0	0.0	86902.0
age	150000.0	86.0	49.0	3837.0
NumberOfOpenCreditLinesAndLoans	150000.0	58.0	6.0	13614.0

3.3. Preprocesamiento

En la parte del preprocesamiento de datos primero particionamos los datos en dos, *train* el cual representan el 70 % de los datos y son utilizados para entrenar el modelo y *test* que representa el restante 30 % de los datos y son utilizados para medir el desempeño del modelo en datos no vistos.

La partición de datos fue realizada utilizando un muestreo estratificado usando como referencia la variable objetivo *SeriousDlqin2yrs* para preservar la distribución de los originales en los datos de *train* y *test*. Además de usar la semilla aleatoria 24 para hacer los resultados reproducibles.

Las variables continuas fueron discretizadas en dos etapas, en la primer etapa utilizamos el método de clasificación fina (fine classing), en esta etapa discretizamos las variables en 20 bins de manera que cada bin contenga aproximadamente el 5 % de información, además consideramos los valores faltantes en una categoría aparte llamada missing.

En la segunda etapa llevamos a cabo una clasificación gruesa (coarse classing), en la cual los resultados de la clasificación fina fueron agrupados en un menor número de bins, uniendo los bins con un nivel de riesgo similar, esto fue llevado a cabo comparando el WoE de los bins.

Para las variables discretas se llevó a cabo el mismo procedimiento, sin embargo, para estas variables no fue necesario llevar a cabo la clasificación fina y solo se aplicó la clasificación gruesa.

Finalmente las variables discretizadas resultantes fueron convertidas a variables numéricas utilizando la transformación WoE para poder ser utilizadas en el modelo logístico.

3.4. Selección de variables

Adicionalmente realizamos un proceso de selección de variables para seleccionar las mejores variables para ser usadas en el modelo. Este proceso fue realizado basándonos en el information value IV de las variables predictoras, seleccionamos como las mejores variables a aquellas que tengan un poder predictivo moderado o fuerte usando como referencia un $IV > 0{,}1$.

Las mejores variables según esta última condición son:

Variable	**IV**
disc_RevolvingUtilizationOfUnsecuredLines	1.1144
disc_NumberOfTimes90DaysLate	0.8856
disc_NumberOfTime30-59DaysPastDueNotWorse	0.7609
disc_NumberOfTime60-89DaysPastDueNotWorse	0.6105
disc_age	0.2267

4. Resultados

Después de haber realizado el preprocesamiento de los datos, seleccionar las mejores variables y de haber realizado la transformación WoE, entrenamos un modelo logístico con los datos *train*.

El entrenamiento fue realizado utilizando el método *sklearn.model_selection.GridSearchCV*, el cual entrena un modelo para cada combinación de hiperparametros y da como resultado el modelo con mejor desempeño (en nuestro caso ocupamos como métrica de desempeño el área bajo la curva ROC). Los hiperparametros usados para entrenar el modelo fueron:

Hiperparametro	**Valores**	**Significado**
C	valores de 0 a 1 con un salto de 0.05	Inverso de la fuerza de regularización
penalty	l_1, l_2	Tipo de regularización
class_weight	None,balanced	Pesos asociados a las clases

Con estos hiperparametros el mejor modelo resultó ser una regresión logística con una regularización l_2, un parámetro C de 0,7 y utilizando los pesos con el valor balanced.Los coeficientes del modelo son:

Variable	**Valor**
intercepto	0.047
disc_RevolvingUtilizationOfUnsecuredLines	-0.668
disc_NumberOfTimes90DaysLate	-0.617
disc_NumberOfTime30-59DaysPastDueNotWorse	-0.625
disc_NumberOfTime60-89DaysPastDueNotWorse	-0.507
disc_age	-0.484

Un aspecto a resaltar es que todos los coeficientes son negativos, esto se debe a que como estamos utilizando la transformación WoE de las variables explicativas y al aumentar el WoE disminuye el riesgo, por lo que la probabilidad de ser malo disminuye.

Después de haber entrenado el modelo, medimos el desempeño del modelo en los datos *train*, los cuales fueron usados para su creación, también medimos el desempeño del modelo en los datos *test* para medir su habilidad de generalización. El desempeño del modelo fue medido usando dos métricas: el área bajo la curva ROC (AUC) y el valor del estadístico K-S.

Datos	**AUC-ROC**	**K-S**
train	0.8504	0.546
test	0.8589	0.560

El desempeño del modelo es bueno, ya que el valor $AUC - ROC$ está cerca de 1, además podemos observar que el modelo es capaz de generalizar, dado

que su desempeño en los datos *train* y *test* es bastante parecido. Por lo que el modelo puede ser utilizado en nuevos datos.

Finalmente, utilizamos el modelo para predecir la probabilidad de que alguien experimentará dificultades financieras en los próximos dos años utilizando un nuevo conjunto de datos. Estas predicciones fueron subidas a la competencia de kaggle: *Give Me Some Credit* https://www.kaggle.com/competitions/GiveMeSomeCredit.

El resultado de nuestro modelo fue un AUC de 0,85242, el cual es parecido al desempeño calculado previamente. Con todo lo anterior consideramos que hemos construido un buen modelo, comparando nuestro resultado con el resultado ganador de la competencia y mejor desempeño (AUC de 0,86955) observamos una diferencia de 0,01713.

5. Segmentación

Referencias

[1] Baesens, B., Roesch, D., Scheule, H. (2016). Credit risk analytics: Measurement techniques, applications, and examples in SAS. John Wiley Sons.

[2] Bolton, C. (2010). Logistic regression and its application in credit scoring (Doctoral dissertation, University of Pretoria).

[3] Han, J., Pei, J., Kamber, M. (2011). Data Mining: Concepts and Techniques. Países Bajos: Elsevier Science.

[4] Hand, D.J. and Henley, W.E. (1997), Statistical Classification Methods in Consumer Credit Scoring: a Review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160: 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

[5] Henley, William Edward (1995). Statistical aspects of credit scoring. PhD thesis The Open University

[6] Laborda, J., Ryoo, S. (2021). Feature Selection in a Credit Scoring Model. Mathematics, 9(7), 746. MDPI AG. Retrieved from http://dx.doi.org/10.3390/math9070746

[7] Orallo, J. H., Quintana, M. J. R., Ramírez, C. F. (2004). Introducción a la Minería de Datos. Pearson Educación.

[8] Siddiqi, N. (2012). Credit risk scorecards: developing and implementing intelligent credit scoring (Vol. 3). John Wiley Sons.

[9] Thomas, L., Crook, J., Edelman, D. (2017). Credit scoring and its applications. Society for industrial and Applied Mathematics.

[10] Zhou, Z. H. (2021). Machine learning. Springer Nature.