Credit scoring

Alejandro Sanchez Madrigal

1 Introducción

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.

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 scorecards o 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.

2 El modelo matemático

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 y 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.

Dentro de los modelos más comunes para construir credit scoring se encuentran:

- Análisis de discriminante
- Árboles de decisión
- K-Nearest Neighbour
- Maquina vector soporte
- Naive Bayes
- Redes neuronales
- Regresión logística

3 Preprocesamiento, preparación y análisis exploratorio de datos

Los datos son la materia prima necesaria para construir un modelo de credit scoring, en ocasiones los datos pueden estar sucios dadas inconsistencias como lo son valores faltantes, valores duplicados, tener un mal tipo de dato, entre otros. Además, pueden presentar propiedades estadísticas no deseadas que pueden afectar el funcionamiento del modelo como lo son los valores atípicos (outliers) o multicolinealidad. Por lo que es necesario aplicar un proceso de exploración ,limpieza y de reducción de variables para poder trabajar con un conjunto de datos limpio, más pequeño y manejable.

3.1 Tipos de datos

Antes de empezar con un análisis, siempre es necesario verificar los tipos de datos de las variables, los principales tipos de datos usados para describir estas variables son:

- Datos continuos: Los elementos se definen en un intervalo que puede ser limitado o ilimitado.
- Datos ordinales: Los elementos están limitados en un conjunto finito con un orden.
- Datos nominales: Los elementos están limitados en un conjunto finito sin un orden.

3.2 Categorización

Uno de los aspectos mas importantes al trabajar con datos, es el tipo de dato de las variables ya que estos indican como deben ser tratados por el modelo. En ocasiones puede ser conveniente reducir el número de categorías, convertir un dato numérico a nominal (discretización) o viceversa (numerización).

- Categorización de variables categóricas: La categorización de variables categóricas es utilizada para reducir el número de categorías de una variable con alta cardinalidad.
- Discretización (binning): Es la conversión de un valor numérico en un valor categórico. La discretización más sencilla consiste en realiza intervalos del mismo tamaño y utilizando el mínimo y el máximo de la variable que queremos discretizar. Para ello se restan el máximo y el mínimo, y el valor resultante se divide por el número de intervalos deseados.
- Numerización: La numerización es el proceso inverso a la discretización, es útil cuando el modelo estadístico que vamos a utilizar no admite datos categoricos. Lo que se suele hacer es lo que se denomina numerización 1 a n, que es la creación de varias variables indicadoras o dummy. Si una variable categórica tiene posibles valores $\{a_1, a_2, ..., a_n\}$ creamos n variables numéricas, con valores 0 o 1 dependiendo de si la variable nominal toma ese valor o no.

3.3 Valores faltantes

La presencia de datos faltantes (missing values) puede ser un problema que puede conducir a resultados poco precisos.Los valores faltantes pueden ocurrir por varias razones y siempre es necesario reflexionar primero sobre el significado de los valores faltantes antes de tomar alguna decisión sobre cómo tratarlos. Las formas más populares para tratar los datos faltantes son los siguientes:

- Eliminar: Consiste en eliminar las observaciones o variables con muchos valores faltantes.
- Ignorarlos: Los valores faltantes pueden ser significativos, además, algunos modelos pueden manejar valores faltantes.
- Remplazar (imputar): Implica remplazar el valor faltante con un valor estimado, tratando de alterar lo menos posible la distribución de la variable.

3.4 Valores atípicos

Los valores atípicos (outliers) son observaciones que son diferentes al resto de la población, estos pueden representar errores en los datos o pueden ser datos correctos diferentes a los demás. Los dos pasos mas importantes para tratar con valores atípicos son la detección y tratamiento.Para poder detectar valores atípicos podemos calcular los percentiles de la distribución o ayudarnos de herramientas visuales como los histogramas o diagramas de cajas. Los tratamientos son muy similares usados a los usados para datos faltantes.

- Eliminar: Eliminar las observaciones con datos atípicos dado que pueden sesgar los datos.
- Ignorarlos:Algunos modelos pueden manejar valores atípicos.
- Remplazar (imputar): Implica remplazar el valor atípicos con un valor estimado, tratando de alterar lo menos posible la distribución de la variable.

3.5 Reducción y análisis de variables

Cuando tenemos un gran número de características en un conjunto de datos, estas pueden estar altamente correlacionadas, esto puede dar origen al problema de multicolinealidad (una situación en la cual existe una relación lineal exacta o aproximadamente exacta entre las variables), o puden haber características irrelevantes, lo que puede provocar que el modelo sobreajuste los datos del conjunto de entrenamiento, es decir, que el modelo capture los patrones en el conjunto de entrenamiento y no sea capaz de generalizar (el cual es uno de los objetivos principales de los modelos de credit scoring) en los datos no vistos. Por lo que es necesario hacer uso del análisis y reducción de variables para tratar de solucionar este problema.

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.

3.5.1 Análisis de componentes principales

El análisis de componentes principales PCA es uno de los métodos más conocidos y utilizados para disminuir la dimensión de un conjunto de datos, tiene por objeto transformar un conjunto de variables $x^1, x^2, \ldots, x^d$ y n observaciones denotado $\mathbf{X}$ en un nuevo conjunto $z^1, z^2, \ldots, z^p$ con $p \leq d$ construidas como combinaciones lineales de las originales, denotado $\mathbf{Z}$. Donde las nuevas variables se generan de manera que sean independientes entre sí y se ordenan de acuerdo con la cantidad de información (varianza) que llevan incorporada. Esto permite

seleccionar las d' primeras variables, asegurándonos que si ignoramos las últimas $p - d'$ variables, estaremos descartando la información menos relevante.

Los pasos para llevar a cabo el análisis de componentes principales son los siguientes:

- Centrar las variables $x^1, x^2, \ldots, x^d$
- Calcular la matriz de covarianzas $\mathbf{X}^T\mathbf{X}$
- Calcular los eigenvalores y los eigenvectores de $\mathbf{X}^T\mathbf{X}$
- Seleccionar los eigenvectores correspondientes a los d' eigenvalores más grandes

El resultado anterior es un nuevo conjunto de variables $z^1, z^2, \ldots, z^{d'}$ con las propiedades anteriores.

El análisis de componentes principales nos permite trabajar con un conjunto de datos con una menor dimensión que el conjunto de datos original. Además, de solucionar el problema de multicolinealidad, cuando proyectamos en una dimensión $d' \leq 3$ podemos visualizar el nuevo conjunto de datos. El principal inconveniente del análisis de componentes principales es que, al crear las nuevas variables como combinación lineal de las variables originales, estas nuevas variables son difícil de interpretar y el modelo pierde explicabilidad.

3.5.2 Clustering de variables

El agrupamiento o clustering consiste en obtener grupos "naturales" a partir de los datos, su objetivo es dividir un conjunto de datos en grupos con características similares, maximizando la similitud de los elementos dentro de un grupo a la vez que minimizamos la similitud con los elementos de otros grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo.

El clustering de variables no divide un conjunto de datos, mas bien, divide un conjunto de variables con características similares utilizando un conjunto de datos. El procedimiento comienza con todas las variables dentro de un cluster y recursivamente elige un cluster que divide en dos sub-clusters, el cluster se elige con base en el porcentaje de variación más pequeño explicado por su componente de cluster. El cluster elegido se divide en dos cluster encontrando los dos primeros componentes principales y asignando cada variable al componente con el que tiene la mayor correlación.

Seleccionamos la variable (o las variables) con el menor $1 - R^2_{ratio}$ como los representantes del cluster.

$$1 - R^2_{ratio} = \frac{1 - R^2_{own}}{1 - R^2_{nearest}}$$

Queremos que las variables representantes del cluster estén lo mas posiblemente correlacionadas con las variables dentro del mismo cluster $R^2_{own} \approx 1$ y lo

menos correlacionadas con las variables en el cluster mas cercano $R^2_{nearest} \approx 0$ Por lo que los mejores representantes del cluster son las variables con el menor $1 - R^2_{ratio}$

3.5.3 Análisis de variables

El análisis de variables involucra analizar cada variable predictora, filtrando las variables más débiles o ilógicas, las variables más fuertes son agrupadas para usarse posteriormente en el modelo.

La fortaleza de las variables es medida usando el poder predictivo de cada atributo de cada variable, generalmente usando el WOE (weight of evidence) o usando el poder predictivo de cada variable usando el IV (Information Value).

3.5.4 WOE

El WOE mide la fortaleza de cada atributo en la separación de clientes buenos y malos, es una medida de la diferencia entre la proporción de buenos y malos en cada atributo. Para una variable categórica con r atributos, sean g_i y b_i el número de buenos y malos en el atributo i, el número de buenos y malos en la muestra son $G = \sum_{i=1}^{r} g_i$ y $B = \sum_{i=1}^{r} b_i$ respectivamente. Una representación cuantitativa del atributo i de la variable está dada por $\ln(\frac{g_i * B}{b_i * G})$

Los valores de la variable predictora son remplazados por el WOE donde el WOE del atributo j de la variable i es dado por:

$$w_{ij} = \ln(\frac{p_{ij}}{q_{ij}})$$

Donde p_{ij} es el número de buenos en el atributo j de la variable i dividido por el número total de buenos y q_{ij} es el número de malos en el atributo j de la variable i dividido por el número total de malos.

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

3.5.5 IV

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

$$IV_i = \sum_{j=1}^{r} (p_{ij} - q_{ij}) * w_{ij}$$

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.

3.6 Estandarización

La estandarización de datos tiene el objetivo de transformar las variables numericas a un rango similar. Algunos de los modelos estadísticos no requieren de estandarización, ya que, si se estandarizan los datos previamente, los resultados pueden ser más difíciles de interpretar. Sin embargo, otros modelos requieren de datos estandarizado para su optimo funcionamiento. Las técnicas de estandarización más comunes son:

- Normalización o z-score: Consiste en restarle la media a la variable y dividir por su desviación estándar, lo que produce una variable con media cero y varianza uno.

$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$

- Estandarización mínimo-máximo: Consiste en transformar una variable tal que esta se encuentre en una escala entre cero y uno.

$$x'_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)}$$

4 Modelos de credit scoring

Existen varios modelos estadísticos que nos permiten construir modelos de credit scoring , todos ellos implican establecer y cuantificar la relación entre las características del solicitante $x_i = (x_{i1}, x_{i2}, ..., x_{id})$ y su desempeño G o B.

References

[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] 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

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

[5] Jorion, Philippe. (2002). Valor en riesgo. México. Limusa Noriega Editores

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

[7] Peña, D. (2002). Análisis de datos multivariantes (Vol. 24). Madrid: McGraw-hill.

[8] Sanche, R., Lonergan, K. (2006, March). Variable reduction for predictive modeling with clustering. In Casualty Actuarial Society Forum (pp. 89-100). Citeseer.

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

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

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