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 atributos (o 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
- Árbol 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 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 ordinal, el orden del atributo ordinal puede ser preservado y utilizado por los pasos subsiguientes o bien puede tratarse el atributo como un valor nominal.
- Numerización: La numerización es el proceso inverso a la discretización, es útil cuando el modelo estadístico que vamos a utilizar no admite datos categóricos.

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 conocido, 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 cuartiles 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.
- Algunos modelos pueden manejar valores atípicos.
- Remplazar (imputar): Implica remplazar el valor atípicos con un valor conocido, tratando de alterar lo menos posible la distribución de la variable.

3.5 Reducción y análisis de variables

Tener un gran número de características en un conjunto de datos 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.

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