Credit scoring

Alejandro Sanchez Madrigal

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.

Estos modelos tienen varias ventajas sobre los modelos tradicionales de evaluación de riesgos, que se basaban en el juicio humano, de las cuales destacan la velocidad de toma de decisiones, la precisión de las decisiones, permite analizar más variables, la reproductibilidad de los modelos, eliminan la subjetividad ,y permite explicar mejor los riesgos que se están tomando.

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.

1.1 Regulación

Los bancos y el sistema financiero en general son un componente importante en la economía mundial, por lo que es necesario que estén altamente regulados por un conjunto de principios y practicas que ayuden a minimizar sus riesgos. Estos principios son conocidos como los acuerdos de Basilea, los cuales representaron un gran avance para la regulación del sistema financiero mundial. Su objetivo es mejorar la solides y seguridad del sistema financiero a través de una mejor administración de riesgos.

La mayor parte de los activos de un banco consisten en portafolios de préstamos, por lo que el riesgo de crédito es de vital importancia para los bancos y sus reguladores. La regulación requiere de una estimación confiable de la distribución de perdidas para cada tipo de portafolio de préstamos. Los acuerdos de Basilea establecen los requerimientos mínimos de capital que deben cumplir los bancos comerciales para realizar coberturas contra el riesgo de crédito, requieren una estimación de la perdida esperada de la forma:

$$EL = PD \times LGD \times EAD$$

Donde:

- PD Probability of default, es la probabilidad de incumplimiento (que no se page el préstamo).
- LGD Loss given default, es la fracción del monto expuesto que no se recuperada en caso de incumplimiento.
- EAD Exposure at default, es una estimación de la medida en que un banco puede estar expuesto en caso del incumplimiento de esa contraparte.
- EL Expected loss, es el costo de proporcionar crédito.

El riesgo de crédito no es el riesgo de pérdidas esperadas del crédito sino el riesgo de pérdidas inesperadas del crédito (volatilidad de las tasas de pérdida reales que ocurren alrededor de EL).El propósito básico del análisis del riesgo de crédito como parte de los acuerdos de Basilea es proporcionar un capital adecuado como red de seguridad contra un posible incumplimiento.

1.2 Credit scoring

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.

En la práctica, el credit scoring se refiere a un problema de clasificación donde un nuevo solicitante de crédito debe ser categorizado en una clase predefinida (típicamente buenos y malos) dependiendo de que tan probable es que incumpla con el pago.

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.

1.3 scorecard

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.

Un tipo especial de puntaje es el puntaje de probabilidades logarítmicas, definido como:

$$S(\mathbf{x}) = \ln\left(\frac{P(G \mid \mathbf{x})}{P(B \mid \mathbf{x})}\right)$$

Aplicando el teorema de Bayes obtenemos:

$$S(\mathbf{x}) = \ln\left(\frac{P(\mathbf{x} \mid G)p_G}{P(\mathbf{x} \mid B)p_B}\right) = \ln\left(\frac{p_G}{p_B}\right) + \ln\left(\frac{P(\mathbf{x} \mid G)}{P(\mathbf{x} \mid B)}\right)$$

Donde p_G y p_B son las proporciones de buenos y malos en la muestra, $P(G|\mathbf{x})$ y $P(B|\mathbf{x})$ son las probabilidades de que un solicitante sea bueno o malo dadas sus características $\mathbf{x}$, $P(\mathbf{x}|G)$ y $P(\mathbf{x}|B)$ son las probabilidades de que un solicitante tenga características $\mathbf{x}$ dado que son buenos o malos respectivamente.

El termino $\ln\left(\frac{p_G}{p_B}\right)$ es el logaritmo de los momios de la muestra, $\ln\left(\frac{P(\mathbf{x}|G)}{P(\mathbf{x}|B)}\right)$ son pesos de evidencia (WOE weights of evidence). Si no tenemos ninguna información sobre las características del solicitante, obtenemos una puntuación igual a los momios de la muestra, conforme conocemos las características del solicitante, el puntaje se va actualizando por el termino que representa el WOE.

1.4 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 Preprocesamiento, preparación y análisis exploratorio de datos

Los datos son la materia prima necesaria para construir un modelo de credit scoring,los datos usados deben estar limpios y ser confiables, pero 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.

2.1 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.

2.2 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.

2.3 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.

2.4 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.

2.5 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.

2.5.1 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 objetivo 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.

2.5.2 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 $1 - R^2_{ratio}$

2.5.3 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).

2.5.4 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.

2.5.5 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.

2.6 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.

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

2.7 Remuestreo

Una muestra con clases no balanceadas ocurre frecuentemente en credit scoring cuando el número de clientes buenos supera ampliamente el número de clientes malos (usualmente el número de cliente malos ronda el 10% del total del conjunto de datos). Este fenómeno puede influenciar el desempeño de los modelos desarrollados.

Las estrategias más populares para solucionar este problema consisten en remuestrear los datos para obtener una distribución de clases alterada. Esto puede ser llevado a cabo, ya sea, sobremuestrando (over-sampling) la clase menos representada (generalmente la clase de malos) o submuestrando (under-sampling) la clase más representada (generalmente la clase de buenos) hasta que ambas clases estén aproximadamente igual representadas.

2.7.1 Over-Sampling

Esta estrategia consiste en expandir la clase menos representada mediante un sobremuestreo aleatorio, a través de una replicación aleatoria de los ejemplos menos representados. El principal inconveniente de este método es que, puede aumentar la posibilidad de sobreajuste, dado que crea copias exactas de la clase menos representada, que no proveen más información para el modelo.

La técnica Synthetic Minority Over-sampling TEchnique (SMOTE) intenta solucionar el problema del sobreajuste, mediante el aumento de ejemplos la clase menos representada, generados artificialmente mediante la interpolación de ejemplos positivos que se encuentran cerca. El proceso para crear los nuevos ejemplos consiste elegir un ejemplo de la clase menos representada, encontrar sus k vecinos más cercanos (usualmente k=5), uno de los vecinos es seleccionado aleatoriamente y el nuevo ejemplo es creado eligiendo un punto entre los dos ejemplos.

2.7.2 Under-Sampling

El submuestreo ayuda a balancear los datos a través de eliminar aleatoriamente datos de la clase mas representada, a pesar de su simplicidad, este método ha mostrado ser uno de los métodos de remuestreo más eficientes. El principal inconveniente de este método es que, puede eliminar datos potencialmente importantes para el proceso de clasificación.

Otros métodos para remover valores de la clase mas representada han sido diseñados, uno de los mas conocidos es el método One-Sided Selection technique (OSS), el cual remueve ejemplos de la clase mas representada que resultan ser redundantes o ruidosos (ejemplos de la clase más representada que rodean los ejemplos de la clase menos representada).

3 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

4 Modelos lineales

Sean $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$ una muestra descrita por d variables. Un modelo lineal intenta aprender una función a través de una combinación de las variables de entrada.

$$f(\mathbf{x}) = b + w_1 x_1 + w_2 x_2 + ... + w_d x_d = b + \mathbf{w}^T \mathbf{x}$$

Donde b y $\mathbf{w} = (w_1, w_2, \ldots, w_d)^T$ son los parámetros del modelo que se necesitan aprender (estimar) para poder hacer predicciones. A pesar de su simplicidad, los modelos lineales son ampliamente usados en la practica y son la base de varios modelos estadísticos y de machine learning, varios modelos no lineales se pueden derivar de los modelos lineales.

4.1 Regresión lineal

Dado un conjunto de datos $\mathcal{D} = \{(x_i, y_i) | \, i = 1, 2, ..m\}$ donde cada x_i es descrito por d variables, es decir, $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ y sea $y_i \in \mathbb{R}$, buscamos estimar los parámetros de un modelo lineal tal que:

$$\hat{Y} = f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \approx Y$$

4.2 Regresión logística

4.3 Análisis de discriminante

5 Evaluación del modelo

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.

5.1 Cross-Validation

La validación cruzada (Cross-Validation) es uno de los métodos mas usados para estimar el error de generalización. Consiste en dividir el conjunto D en k subconjuntos disjuntos tales que $D = D_1 \cup D_2 \cup \ldots \cup D_k$ y $D_i \cap D_j = \varnothing$ si $i \neq j$.

En cada una de las k iteraciones utilizamos $k-1$ subconjuntos para entrenar el modelo y el subconjunto restante es utilizado para estimar el error de generalización, usando exactamente una sola vez cada subconjunto como conjunto de prueba. Promediamos los resultados de las k pruebas para estimar el error de generalización.

5.2 Medidas de desempeño

Para evaluar la habilidad de generalización de un modelo, es necesario definir una medida de desempeño que nos permita cuantificar la habilidad de generalización. En las tareas de predicción evaluamos el desempeño de modelo comparando los valores predichos por el modelo $\hat{y}$ con el valor real y.

5.2.1 Mean Squared Error

El método mas usado para medir el desempeño de un modelo en una tarea de regresión (cuando $y \in \mathbb{R}$) es el error cuadrático medio MSE definido como:

$$MSE = E(f; D) = \frac{1}{m} \sum_{i=1}^{m} \left(f\left(x_i\right) - y_i \right)^2 = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}_i - y_i \right)^2$$

5.2.2 Tasa de error y exactitud

La tasa de error y exactitud son medidas de desempeño para tareas de clasificación. La tasa de error es la proporción de muestras mal clasificadas con respecto al tamaño total de las muestras y está definida como:

$$E(f; D) = \frac{1}{m} \sum_{i=1}^{m} I\left(f\left(x_i\right) \neq y_i \right) = \frac{1}{m} \sum_{i=1}^{m} I\left(\hat{y}_i \neq y_i \right)$$

La exactitud (accuracy) es la proporción de muestras correctamente clasificadas con respecto al tamaño total de las muestras y está definida como:

$$acc(f; D) = \frac{1}{m} \sum_{i=1}^{m} I\left(f\left(x_i\right) = y_i \right) = \frac{1}{m} \sum_{i=1}^{m} I\left(\hat{y}_i = y_i \right) = 1 - E(f; D)$$

5.2.3 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 $P = \frac{TP}{TP+FP} = \frac{TP}{P'}$.

Para evaluar completamente la efectividad de un modelo, debes examinar la precisión y la recuperación, pero estas medidas son contradictorias, cuando la precisión es alta la recuperación es baja y viceversa.

5.2.4 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

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.

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] García, V., Marqués, A.I., Sánchez, J.S. (2012). Improving Risk Predictions by Preprocessing Imbalanced Credit Data. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds) Neural Information Processing. ICONIP 2012. Lecture Notes in Computer Science, vol 7664. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34481-7-9

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

[5] 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

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

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

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

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

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

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

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

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

https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419