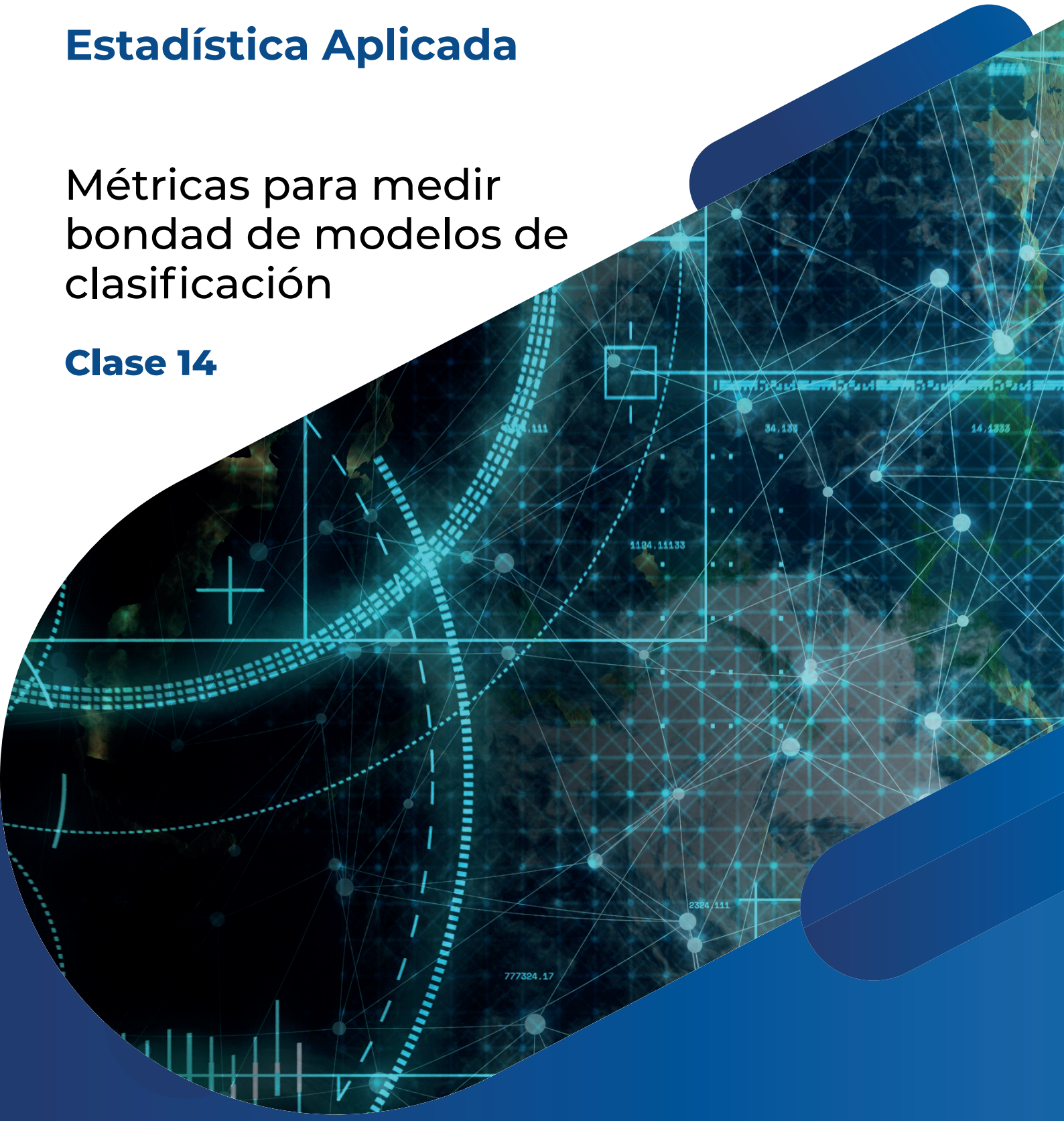
Estadística Aplicada

Métricas para medir bondad de modelos de clasificación

Clase 14



Ingeniería en ciberseguridad

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

En la presente clase revisaremos las distintas métricas que sirven para evaluar los modelos logísticos. Como veremos, existen varias métricas, y su uso y aplicabilidad dependerán de algunos factores y del propósito general de clasificación.

Un tema de particular importancia será el tratamiento de *la data* cuando las clases se hallan desbalanceadas, explicaremos los resultados de la métrica cuando esto ocurre, así como los mecanismos para compensar el desbalanceo.

Clase 14:

Resultado o resultados de aprendizaje que será abordado con el contenido de la clase.

Analizar información contextual sobre hábitat, infraestructura y movilidad, mediante técnicas estadísticas descriptivas e inferenciales, para la adecuada toma de decisiones.

Reto # 4

14) Métricas para medir bondad de modelos de clasificación

Antes de entrar en el análisis de las métricas para evaluar la bondad de los modelos es importante saber interpretar los coeficientes de las variables en este tipo de modelos.

Un tópico que llama la atención en regresiones logísticas es interpretar los coeficientes del modelo de regresión, debido a que están en la escala log-odds. Debemos tener cuidado de convertirlos antes de interpretar los términos de las variables originales, recordando la definición de logit.

Figura 1

Fórmula de logit

Creación del autor Alfonso Prado

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1$$

Entonces la función exp(β0) representa las probabilidades de que la característica de éxito esté presente para un individuo x=0, es decir en la línea de base. Si hay múltiples predictores involucrados, todos deberían establecerse en 0 para esta interpretación.

La función exp(βi) representa el aumento multiplicativo en las probabilidades de éxito por cada aumento de 1 unidad en x. Si βi > 0 indica que es una relación positiva, si es negativo entonces el incremento de X decrementa la posibilidad de que se presente el evento.

Por ejemplo, de acuerdo con el modelo creado en la clase anterior el coeficiente de la variable "MaritalStatus.Single " es 0.8421 , y su exp(0.8421) es 2.32. Esto debe interpretarse de la siguiente

forma: Por cada unidad que aumenta la variable MaritalStatus.Single el **odds** (no el OR) de que se presente la renuncia aumenta en 2.32.

Figura 2

Modelo creado en la clase anterior

Creación de autor Alfonso Prado

```
Call:  glm(formula = Attrition ~ MaritalStatus.Divorced + MaritalStatus.Single +
    MaritalStatus.Married, family = binomial(link = "logit"),
    data = data_dv)

Coefficients:
            (Intercept)  MaritalStatus.Divorced    MaritalStatus.Single
                -1.9235                 -0.2843                   0.8421
 MaritalStatus.Married
                     NA

Degrees of Freedom: 4299 Total (i.e. Null);  4297 Residual
Null Deviance:      3804
Residual Deviance: 3680         AIC: 3686
```

14.1) Matriz de confusión

Es importante mencionar que, para optimizar el rendimiento de cualquier modelo, deben primero realizarse actividades de limpieza de datos y preprocesamiento, lo que incluye la gestión de datos atípicos, transformaciones, normalización, entre otros.

Luego nos preguntamos ¿cómo podemos medir la efectividad de nuestro modelo en términos de eficiencia, rendimiento y precisión? Ahí entra **la matriz Confusión.**

De acuerdo con Hilbe (2015), esta matriz es la medida del rendimiento para la clasificación y es ampliamente utilizada en el aprendizaje automático(ML).

La matriz de confusión es un tipo especial de tabla de contingencia, con dos dimensiones (“real” y “predicha”) y conjuntos idénticos de “clases” en ambas dimensiones (cada combinación de dimensión y clase es una variable en la tabla de confusión). El nombre proviene del hecho de que facilita ver si el sistema está confundiendo dos clases (es decir, etiquetando erróneamente una como otra). Generalmente, las tablas contienen la predicción en las filas y los valores actuales (observaciones) en las columnas, aunque hay autores que lo colocan al revés.

Figura 3

Matriz de Confusión

Creación del autor: Alfonso Prado

En la tabla vamos a identificar entonces 4 valores:

- Verdaderos positivos: Identificados como TP predicción positiva y, en realidad, es positivo.
- Falsos positivos: Identificados como FP predicción positiva y, en realidad, es negativo.
- Falsos negativos: Identificados como FN predicción negativa y, en realidad, es positivo.
- Verdaderos negativos: Identificados como TN predicción negativo y, en realidad, es negativo.

14.2 Curvas Características de la Separabilidad

Basados en la matriz de confusión, podemos establecer algunos indicadores de la eficiencia del modelo:

Accuracy: Simplemente mide con qué frecuencia el clasificador hace la predicción correcta. Es la relación entre el número de predicciones correctas y el número total de predicciones, se utiliza cuando los verdaderos positivos y los verdaderos negativos son más importantes.

Sensibilidad: Calculada como el número de predicciones positivas correctas dividido por el número total de positivos reales (observaciones). Algunos autores la denominan recuperación (REC) o tasa de verdaderos positivos (TPR).

Especificidad: Calculada como el número de predicciones negativas correctas dividido por el número total de negativos. Algunos autores la denominan tasa negativa verdadera (TNR).

Tasa de Falsos Positivos (FPR): Se define como la relación entre falsos positivos y la suma de falsos positivos más verdaderos negativos. Es el inverso de la especificidad (1-especificidad).

Precisión: Representa la proporción de casos positivos que se predicen correctamente dividido para el total de casos positivos que se predicen. La diferencia entre sensibilidad y precisión radica en que la sensibilidad se calcula en base a los casos reales (observaciones) que son positivos, mientras que la precisión se calcula en base a las predicciones positivas.

Tabla 1

Fórmulas de las métricas derivadas de la tabla de confusión

Creación del autor Alfonso Prado

Accuracy	Sensibilidad (REC o TPR)	Especificidad (TNR)	Falsos Positivos (1- TNR)	Precisión
$\frac{TP + TN}{TP + TN + FP + FN}$	$\frac{TP}{TP + FN}$	$\frac{TN}{TN + FP}$	$\frac{FP}{FP + TN}$	$\frac{TP}{TP + FP}$

La pregunta que nos hacemos es: ¿Cuál de las métricas deberíamos utilizar? Todos los indicadores anteriores sirven para verificar una parte del modelo; por ejemplo, qué tan bien se predicen los TP o TN, y funcionan bien cuando los datos están balanceados. Se considera que los datos están balanceados cuando la cantidad de casos de cada clase es similar, y desbalanceados cuando existen diferencias significativas.

Veamos un ejemplo: Asuma que tenemos 1000 casos positivos de los cuales el 90% son predichos correctamente y 10% incorrectamente. Asuma que tenemos 50 casos negativos de los cuales podemos predecir correctamente el 50%. De la Tabla 2 obtenemos: TPR=900/925 =0.97 =97% y TNR=25/125=0.2 =20% . Notamos que TPR está sobreestimado y TNR esta subestimado, este es el efecto que se obtiene cuando la data no está balanceada.

En estos casos, debemos recurrir a la métrica llamada F1 o a las curvas características de separabilidad, que resumen de mejor manera la eficiencia del modelo.

Figura 4

Fórmula de la métrica F1

Creación del autor Alfonso Prado

$$PuntajeF1 = \frac{2 * Precision * Sensibilidad}{Precision + Sensibilidad}$$

Otras alternativas posibles en caso de data desbalanceada consisten en previamente balancear la misma en forma manual, para luego entrenar el modelo. Existen dos técnicas para balancear conocidas como: Undersampling y Oversampling:

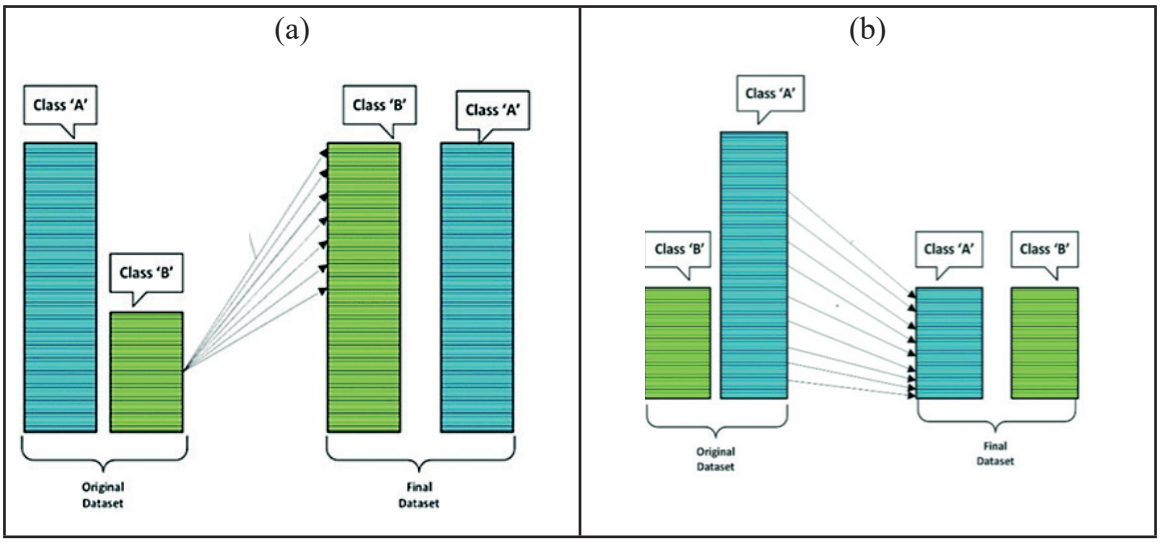
- ✓ El “Under-sampling” (submuestreo) funciona reduciendo el tamaño de la clase mayoritaria para que coincida con la clase menos prevalente.
- ✓ El “over-sampling” (sobre muestreo) se utiliza para aumentar el tamaño de muestras minoritarias. La técnica conocida como SMOTE (Synthetic Minory Over-sampling Technique) ofrece esta posibilidad en el paquete caret.

La siguiente figura explica el concepto

Figura 5

Sobre muestreo (a) y Submuestreo (b)

Creación del autor Alfonso Prado



La implementación del submuestreo es relativamente fácil; implica eliminar una cierta cantidad de observaciones. Pero en sobremuestreo se podría pensar en duplicar las observaciones de dicha clase, lo cual equilibrará la distribución de clases; sin embargo, esto no mejorará el rendimiento del modelo, ya que no le proporciona información adicional.

SMOTE.- Esta técnica selecciona los ejemplos minoritarios que están cerca en el espacio de características. En ML las "características" son sinónimo de variables; por lo tanto, el espacio de características es el espacio de valores de dicha variable. Dentro de este espacio, SMOTE crea una nueva muestra en algún lugar y le asigna el valor del vecino más cercano utilizando el algoritmo de **KNN** (K Nearest Neighbors). En otras palabras, el algoritmo crea un ejemplo aleatorio cuya clase se fija de acuerdo con el K vecino más cercano. Esta observación "sintética" se crea entre dos ejemplos en el espacio de características.

El uso de SMOTE tiene una desventaja, ya que no considera la clase mayoritaria al crear ejemplos sintéticos. En otras palabras, el algoritmo funciona bien cuando las clases están bien separadas; esto puede causar problemas cuando hay una superposición entre las clases.

SMOTE

Presenta detalle para su implementación

Enlace

Veamos un ejemplo de implementación de SMOTE

```
#Validamos si la data esta balanceada

prop.table(table(data$Attrition))

        No        Yes
0.8383721 0.1616279

#Obviamente no está balanceada

#En este caso utilizaremos la función downSample cuya sintaxis es:

#downSample(x, y, list = FALSE, yname = “Class”)

#x=lista de variables predictoras

#y=variable de clases , en este caso Attrition

#list=FLASE indica que es un dataframe

#yname= nombre de la etiqueta de clase


set.seed(100)

dataBalanceada <- caret::downSample(x=data %>% select(-Attrition) ,
                                    y=data$Attrition,
                                    yname=”Attrition”)


prop.table(table(dataBalanceada$Attrition))

  No Yes
 0.5 0.5


#Ahora deberíamos repetir la creación y análisis de la matriz de confusión para asegurar que sensibilidad y especificidad están correctos
```

En última instancia, la mejor métrica es la que le conviene al negocio.

Por ejemplo, suponga que obtenemos una sensibilidad=0,35 y la especifidad=0,90. Por tanto, este modelo mucho más específico que sensible. Esta es la situación que nos interesa cuando nuestro objetivo es evitar a toda costa los falsos positivos, que sería el caso en el cual, como conclusión del estudio, una empresa estuviera injustamente penalizando a un cliente.

Por el contrario, si los falsos positivos no nos preocupan tanto y lo que queremos evitar son los falsos negativos, nos interesa una mayor sensibilidad o **recall**. Por ejemplo, no nos importa un falso positivo en una prueba de diabetes, ya que la prueba se repetirá sin duda. Sin embargo, no nos interesa que una persona diabética no diagnosticada no acceda rápidamente al tratamiento adecuado debido a un falso negativo.

Pero en la práctica, los modelos nunca son perfectos. Para ayudarnos a visualizar la bondad de los modelos recurrimos a la visualización mediante diagramas que permiten comparar dos métricas simultáneamente. Estas curvas permiten visualizar la separabilidad entre clases. Idealmente, se desea que las observaciones positivas sean predichas como positivas y lo mismo con las negativas.

Curva ROC

Una curva ROC es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. Esta curva usa dos métricas en sus ejes: la tasa de verdaderos positivos TPR (sensibilidad) y la tasa de falsos positivos FPR (1-especificidad).

En principio deseamos que la sensibilidad sea lo más grande posible, mientras el FPR sea lo más bajo posible. La gráfica resultante sería la figura 6(a).

En la medida que el modelo empieza a perder su capacidad de discriminar entre clases positivas y negativas, la curva empieza a acercarse a una diagonal. Cuando llega a ser una diagonal es la peor situación, porque indica que el modelo no puede discriminar entre clases y, por lo tanto, no es utilizable.

Cuando la tasa de falsos positivos llega a ser igual a 1, el modelo en realidad está invirtiendo las clases. Significaría que el modelo predice una clase negativa como una clase positiva y viceversa, en cuyo caso la curva ROC se verá como en la figura 7(c). Dado que TPR y FPR tienen un valor máximo de 1, el área bajo la curva tendrá un valor de 1. De acuerdo con Hilbe (2015), “Los valores de 0,5 a 0,65 tienen un poder predictivo bajo. Los valores de 0,65 a 0,80 tienen un valor predictivo moderado. Muchos modelos logísticos se ajustan a este rango. Los valores superiores a 0,8 e inferiores a 0,9 generalmente se consideran de alto poder predictivo”.

La curva se crea basado en un umbral del valor de la variable; para cada umbral escogido se calculan las dos métricas mencionadas y se grafican los puntos, que luego se unen por medio de una curva, dando como resultado la figura siguiente.

Figura 6

Casos extremos de la curva ROC

Creación del autor Alfonso Prado

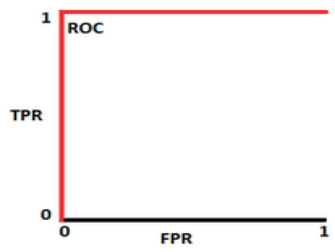
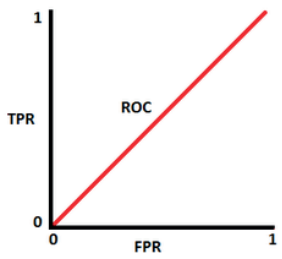
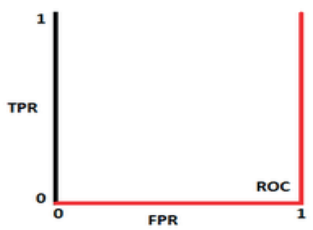
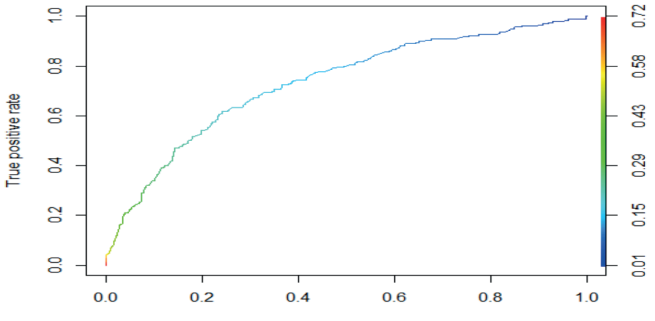
Modelo ideal(a)	Modelo sin capacidad de discriminación(b)	Modelo confunde las clases(c)
		

Figura 7

Curva ROC de un modelo real

Creación del autor Alfonso Prado



A fin de obtener una métrica objetiva se estila calcular el área bajo la curva. De esta manera, pequeñas diferencias entre curvas de distintos modelos son fáciles de notar y comparar. Un modelo excelente tiene AUC cerca de 1, lo que significa que tiene una buena medida de separabilidad entre las clases. Un modelo pobre tiene un AUC cercano a 0.5, lo que significa que tiene la peor medida de separabilidad; esto es, no puede distinguir entre una clase y otra, y no es mejor que clasificar al azar. Medidas menores a 0.5 indicarían que los casos positivos se predicen como negativos y viceversa.

Curva PR sensibilidad vs precisión

La curva PR es la gráfica resultante de relacionar la precisión y la sensibilidad (REC o TPR) de un modelo. Nos permite ver a partir de qué valor de sensibilidad obtenemos una degradación de la precisión.

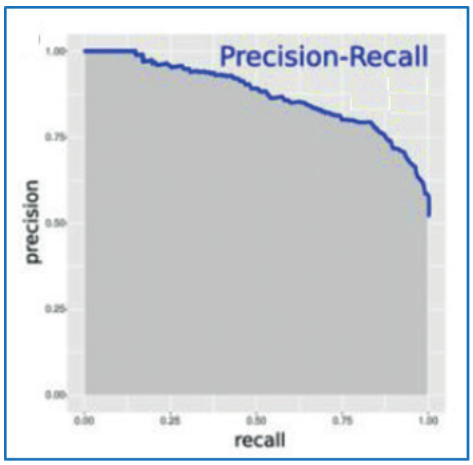
Note que en realidad lo que está comparando es FP y FN, es decir, la diagonal secundaria. En un modelo perfecto sería aquel que pasa por el punto (1,1); por tanto, cuanto más se acerque a esa esquina superior derecha mejor es el comportamiento del modelo. Al igual que con la curva ROC, también se puede calcular el área bajo la curva.

Es preferible usar esta curva en vez de la curva ROC en casos en los que exista un desbalanceo en las clases, ya que la curva ROC puede dar una visión optimista del modelo debido a la dependencia de los falsos negativos que, en conjuntos con clases desbalanceadas, aumentará.

Figura 8

Curva precisión-sensibilidad

Creación del autor Alfonso Prado



Cuándo usar cada curva :

La curva ROC es Ideal para conjuntos de datos equilibrados donde las clases positivas y negativas tienen la misma importancia. Es útil cuando el objetivo es encontrar el equilibrio óptimo entre sensibilidad y especificidad.

La Curva de precisión-sensibilidad es más informativa para conjuntos de datos desequilibrados donde la clase positiva es poco frecuente o más importante. Es útil cuando el objetivo es maximizar la precisión de las predicciones positivas.

Las curvas ROC consideran tanto las clases positivas como las negativas, mientras que las curvas de precisión-sensibilidad se centran en la clase positiva y son más útiles en caso de data desbal-

anceada.

Mejores prácticas en la creación de modelos logísticos

A continuación, presentamos un conjunto de mejores prácticas que puede ayudar a mejorar el rendimiento de un modelo:

- ✓ En la selección de variables elimine primero variables que no demuestran correlación o aquellas numéricas que tengan varianza 0 o cercana a 0.
- ✓ Si existen variables con NAs proceda con las imputaciones que considere necesarias; sin embargo, tome en cuenta que las imputaciones podrían cambiar completamente la forma de los datos
- ✓ Para variables continuas que no presenten una buena distribución normal se puede utilizar una transformación logarítmica. Esto ayudará a que los datos no estén tan dispersos. Sin embargo, note que la transformación debe realizarse antes de cualquier operación de normalización, ya que esto podría llenar el dataset de NAs para los valores 0 o negativos. En este caso deberá utilizar Yeo-Johnson.

  Normalización Yeo-Johnsson

  Explica el funcionamiento de normalización para valores negativos

  Enlace

- ✓ Estandarice los datos y luego aplique one-hot. One-Hot da como resultado que los datos estén más dispersos, lo cual muchos algoritmos pueden usar de manera eficiente. Si estandariza los datos primero, creará datos densos con los cuales los algoritmos se ejecutarán de manera menos eficiente.
- ✓ Lo mismo ocurre con la agrupación de categorías; primero haga los agregados y luego codifique con one-hot.

Referencias citadas en la Clase 14.

Hilbe J., (2015), Practical Guide to Logistic Regression, CRC Press

Rocha A.,(2020) Codificación de variables categóricas en

aprendizaje automático, Universidad de Sevilla

Ramirez-Alan O,(2016), Chi-Cuadrado https://rpubs.com/osoramirez/111403

Mehdi K, Amiri A. (2019) The Effect of Link Function on the Monitoring of Logistic Regression Profiles, https://www.semanticscholar.org/

Definición de los términos citados en la Clase 14.

Odds y log Odds	Las probabilidades (técnicamente, las probabilidades de éxito) se definen como la probabilidad de éxito/probabilidad de fracaso. Por lo tanto, la probabilidad de éxito (80 % de probabilidad de lluvia) tiene una probabilidad de fracaso (20 % de probabilidad de que no llueva); como ecuación (la "razón de probabilidades"), es 0,8/0,2 = 4. El logaritmo de las probabilidades es el logaritmo de las probabilidades. ln(4) = 1,38629436 ≅ 1,386.
Imputación	La imputación de datos es una técnica estadística que se utiliza para reemplazar valores faltantes o nulos en un conjunto de datos. En el análisis de datos y la ciencia de datos, el manejo de los datos faltantes es crucial, ya que pueden afectar significativamente los resultados de cualquier análisis. La imputación de datos tiene como objetivo proporcionar un conjunto de datos más completo, lo que permite un modelado y un análisis más precisos. Existen varios métodos para la imputación de datos, cada uno con sus fortalezas y debilidades, según la naturaleza de los datos y el contexto del análisis. La imputación de media, mediana y moda se encuentran entre las formas más simples de imputación de datos. La imputación de media reemplaza los valores faltantes con el promedio de los datos disponibles, mientras que la imputación de mediana utiliza el valor medio; y la imputación de moda utiliza el valor que aparece con mayor frecuencia. Estos métodos son fáciles de implementar y comprender, pero pueden introducir sesgos, especialmente en distribuciones sesgadas.

Profundización Clase 14.

Matriz de confusión

Presenta casos de uso

Enlace

PUCE Virtual

La excelencia no se improvisa

síguenos