Aprendizaje Estadístico Supervisado

Natalia da Silva

2024

Evaluación de modelos

- Para saber si un modelo es más apropiado que otro para un problema de interés es necesario evaluar su performance en dicho problema.
- No hay un métodos estadísticos mejores a otros para todos los posibles problemas.
- Algunos métodos funcionan bien en algunos problemas y no en otros.
- Hay que decidir para cada problema cuál método produce el mejor resultado.
- La naturaleza del conjunto de datos y del algoritmo es lo que determina al final lo apropiado o no de una técnica para un problema de interés.
- Seleccionar la mejor solución es uno de los puntos más desafiantes del aprendizaje estadístico.

Teoría de la decisión

- Sea $X \in R^p$, vector aleatorio de valores reales
- $Y \in R$ o $Y \in \{1, 2, \ldots G\}$ dependiendo del problema, con distribución $Pr(Y, X)$.
- Buscamos $f(X)$ para *predecir* Y.

Se define una función de pérdida $L(Y, f(X))$ que penaliza los errores de predicción.

Las más utilizadas:

- **Regresión:** pérdida cuadrática: $L(Y, f(X)) = (Y - f(X))^2$
- **Clasificación** pérdida 0-1: $L(Y, f(X)) = I(Y \neq f(X))$

Teoría de la decisión estadística

La teoría de la decisión unifica el tratamiento estadístico de problemas de regresión y clasificación.

Criterio para elegir f: minimizar *error esperado de predicción*:

$$\mathrm{EPE}(\mathrm{f}) = \mathrm{E}\left[\mathrm{L}(\mathrm{Y}, \mathrm{f}(\mathrm{X}))\right]$$

El objetivo es hallar $\mathrm{f}()$ para minimizar la pérdida esperada, $\mathrm{E}\left[\mathrm{L}(\mathrm{Y}, \mathrm{f}(\mathrm{X}))\right]$, que

Puede obtenerse *para cada valor de* x minimizando

$$\mathrm{E}\left[\mathrm{L}(\mathrm{Y}, \mathrm{f}(\mathrm{X}))|\mathrm{X} = \mathrm{x}\right]$$

para cada valor de x

- Cambiar la función de pérdida, $\mathrm{L}()$, modifica el predictor óptimo.

Problema de Regresión con pérdida cuadrática

$$L(Y, f(X)) = (Y - f(X))^2$$

¿Cómo elegimos f según esta función de pérdida?

$$\begin{aligned} EPE(f) &= E(Y - f(X))^2 \\ &= E_X E_{Y/X}[(Y - f(X))^2 / X] \end{aligned}$$

es suficiente minimizar EPE punto a punto

$$f(x) = \arg\min_c E_{Y/x}[(Y - c)^2 / X = x]$$

La solución de esto es $f(x) = E(Y/X = x)$

El mejor predictor de Y es la esperanza condicional en todo punto $X = x$ cuando uso la pérdida cuadrática.

Problema de Regresión con pérdida cuadrática

Con un método concreto como el vecino más cercano para regresión vemos que intenta cumplir con este predictor óptimo utilizando los datos que tiene

Dos pasos:

- La esperanza es aproximada con la media en los datos de la muestra
- En vez de condicionar en un punto lo hace en una región o vecindario cercano al punto objetivo

Clasificación con pérdida 0-1

- Si ahora Y es categórica se utiliza el mismo paradigma simplemente utilizando una función de pérdida diferente para penalizar el error de predicción
- $Y \in \{1, 2 \ldots G\} = G_r$, conjunto de las clases posibles
- La función de pérdida es una matríz L_{GxG}, usando la pérdida 0-1 tiene ceros en la diagonal y 1 fuera de ella

$$\begin{aligned} EPE(f) &= E[L(Y, f(X))] \\ &= E_X E_{Y/X}[L(Y, f(X))/X] \\ &= E_X \sum_{g=1}^{G} L(g, f(X))P(g/X) \end{aligned}$$

Es suficiente minimizar EPE punto a punto

Clasificación con pérdida 0-1

Con pérdida 0-1 $L(g, g^*) = I(g \neq g^*)$

$$
\begin{aligned}
f(x) &= \arg \min_{g^* \in G_r} \sum_{g=1}^{G} L(g, g^*) P(g/X = x) \\
&= \arg \min_{g^* \in G_r} \sum_{g^* \in G_r / g \neq g^*} P(g/X = x) \\
&= \arg \min_{g^* \in G_r} [1 - P(g/X = x)] \\
&= \arg \max_{g^* \in G_r} P(g/X = x)
\end{aligned}
$$

El predictor óptimo de Y con la pérdida 0-1 es el clasificador de Bayes. Se predice la clase más probable.

Clasificación con pérdida 0-1

Vecino más cercano para el problema de clasificación directamente aproxima esta solución

$$\hat{f}(X) = \arg\max_{g \in G_r} \sum_{x_i \in N_k(x)} I(y_i = g)$$

- La probabilidad se estima con una proporción en la muestra
- En vez de condicionar en un punto lo hace en un vecindario

Teoría de la decisión estadística

La teoría de la decisión unifica el tratamiento estadístico de problemas de regresión y clasificación.

Las pérdidas más utilizadas

Problema	**Pérdida**	**Predictor**
Regresión	$L(Y, f) = (Y - f(x))^2$	$E(Y \mid X)$
Clasificación	$L(Y, f) = I(Y \neq f(x))$	$\text{argmax}_{g^*} \Pr(Y = g^* \mid X)$

Muchas pérdidas posibles: valor absoluto, percentil, Huber, entropía.

Error de generalización

EPE no lo podemos calcular

- En la práctica no conocemos $\Pr(Y, X)$
- Disponemos de n datos observados $Tr = (y_i, x_i)_{i=1}^{n}$
- Con los datos se contruye un estimador $\hat{f}()$
- Consideramos un **nuevo** punto (y_o, x_o)

Definimos **error de generalización** (error test) como:

$$Err_{Tr} = E\left[L(\hat{f}(x_o), y_o)|Tr\right]$$

es el error de predicción en *nuevos puntos* (y_o, x_o) condicional en los *datos observados* Tr.

Error de generalización *esperado*

Luego, definimos **error de generalización esperado** como

$$\mathrm{Err} = \mathrm{E}(\mathrm{Err}_{\mathrm{Tt}}) = \mathrm{E}_{\mathrm{Tr}}\left(\mathrm{E}_{(\mathrm{x},\mathrm{y})|\mathrm{Tr}}\left[\mathrm{L}(\mathrm{f}\hat{(}\mathrm{x}_{\mathrm{o}},), \mathrm{y}_{\mathrm{o}})|\mathrm{Tr}\right]\right)$$

donde $\mathrm{E}()$ incorpora todo lo que es aleatorio: el conjunto de datos observado, la distribución conjunta (x, y) y la estimación $\mathrm{f}\hat{()}$.

- También se le llama Error de predicción esperado.
- Si bien el objetivo sería estimar $\mathrm{Err}_{\mathrm{Tr}}$, en general no es posible estimarlo.
- Veremos técnicas para aproximar Err.

Estimar error

¿Se puede usar err (Error de entrenamiento) para estimar Err (Error de Generalización Esperado)?

NO!

- Generalmente el error con los datos de entrenamiento es *optimista* (más pequeño que con datos nuevos) ya que los datos de entrenamiento fueron usados para ajustar el modelo.
- Por diseño el error es más pequeño que el error calculado con nuevos datos.
- Siempre disminuye cuando el modelo vuelve muy complejo, lleva a sobre-ajuste

Evaluación

- En aprendizaje supervisado podemos evaluar la efectividad de un algoritmo basado en el error que comete.
- En problemas de clasificación se calcula el porcentaje de veces que el algoritmo clasifica incorrectamente (error de clasificación) o el porcentaje correctamente predicho (precisión)
- Para problemas de regresión comunmente se compara la diferencia entre el valor predicho y el verdadero valor, el MSE, ….

Evaluación

- Hemos mencionado que para evaluar los modelo es importante separar la muestra en entrenamiento y testeo.
- No usar los mismos datos para entrenar el modelo y evaluarlo.
- Estimar el error de predicción con los datos de entrenamiento subestima el error de generalización esperado Err.

Particionar los datos

- Se deben separar los datos en, datos de entrenamiento y datos de testeo.
- En general se recomienda 2/3 de los datos para entrenar el modelo y el resto (1/3) es usado para testear el modelo.
- Con los datos de entrenamiento ajusto el modelo
- Con los datos de test uso el modelo para predecir estas observaciones que no fueron utilizadas para construir el modelo y calculo errores de predicción.

Como separar en training y test puede tener el problema de variar mucho de muestra a muestra, surgen alternativas.

Técnicas de remuestreo

- Generar (re)-muestras a partir de UN conjunto de datos de entrenamiento.
- Se usan para evaluar propiedades estadísticas: errores estándard, sesgo, error de predicción.
- Selección de parámetros de ajuste basados en remuestreo

Vimos: Validación cruzada y Bootstrap.

Estas técnicas son tan generales que pueden ser utilizadas para la mayoría de los método de aprendizaje

Ejemplo:Abandono 1ero 2016-2017

Modelos de clasificación para el abandono

- Variable de respuesta categórica con dos niveles (abandona, no abandona).
- Variables explicativas: Contexto sociocultural, sexo, extra edad fuerte, extra edad leve, inasistencias relativas, centro educativo, departamento.
- Problema: los datos están muy desbalanceados sólo un 7% abandonan

Performance de clasificadores

Existen muchas medidas para evaluar clasificadores, las más básicas

- Matriz de confusión
- Sensibilidad y Especificidad
- Curva ROC
- Área debajo de la curva

Vemos las medidas en el ejemplo de abandono

Observado vs predicho, matriz de confusión

Matriz de confusión para dos clases (Yes, No)

Las celdas de la diagonal principal tienen las clases que son predichas correctamente por el modelo, mientras que las que están fuera de la diagonal indican errores en cada uno de los casos posibles.

Observado vs predicho, matriz de confusión

Matriz de confusión para dos clases (Yes, No)

Las celdas de true positive (TP) y true negative (TN) contienen el número de casos correctamente predichos como **1** o **0**. Las celdas de false positive (FP) y false negative (FN) representan el número de casos que son predichos incorrectamente de **1** y **0**.

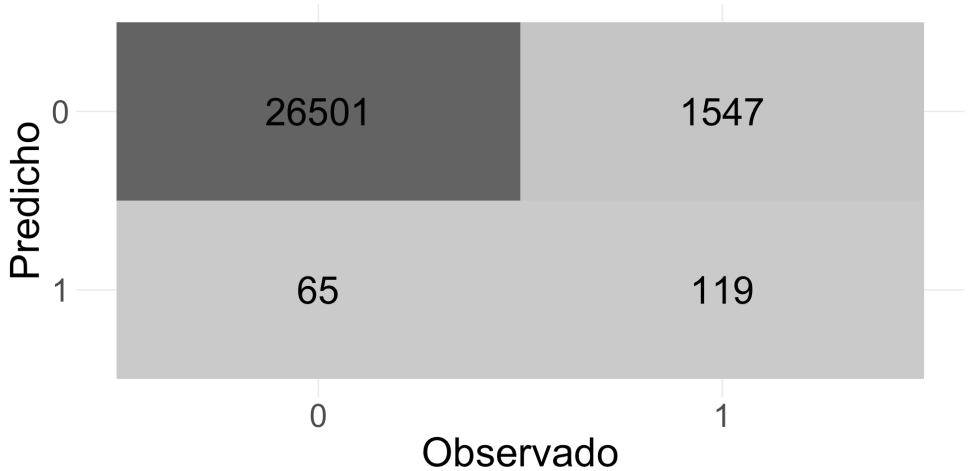
Precisión

- En problemas de clasificación en general se calcula la tasa precisión del modelo, sin embargo no siempre nos da la idea completa sobre la performance del modelo.
- La precisión global es la métrica más común.
- Presentaremos algunas limitaciones de la precisión de predicción y algunas métricas adicionales que nos dan alguna perspectiva adicional en la performance del modelo.

Precision vs error

- Precisión $\frac{(TP+TN)}{(TP+TN+FN+FP)}$
- Error: $\frac{(FN+FP)}{(TP+TN+FN+FP)}$
- Una de las limitaciones de la precisión es que no considera el tipo de error que se comete
- Hay que tener medidas que también permitan discriminar entre clases.
- Podríamos predecir bien en general el no abandono pero no hacerlo bien con el abandono por lo que puede convenir separar las mediadas según sean buenas para predecir clases positivas o negativas dependiendo de lo que nos interesa más para el problema.

Matriz de confusión



- 65 obs son clasificadas erroneamente como 1 (FP)
- 1547 obs son clasificadas erroneamente como 0 (FN)

- Precisión: $\frac{(TP+TN)}{(TP+TN+FN+FP)} = \frac{(26501+119)}{(26501+119+1547+65)} = 0.94$
- Error: $\frac{(FN+FP)}{(TP+TN+FN+FP)} = \frac{(1547+65)}{(26501+119+1547+65)} = 0.057$

Sensibilidad y Especificidad

A partir de la Matriz de confusión se obtienen *medidas de ajuste*.

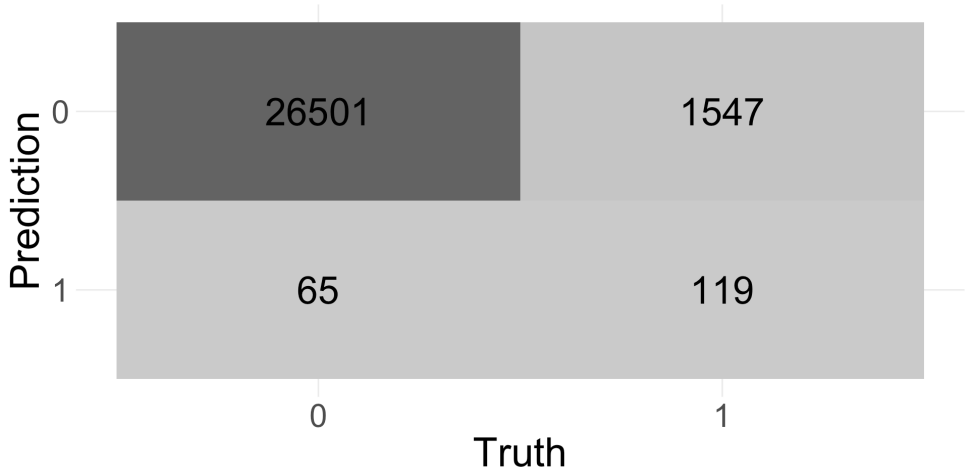
- **Sensibilidad** del modelo, es la proporción de verdaderos positivos (TP) que son correctamente identificados por el modelo (tasa de verdaderos positivos).
- **Especificidad** es la tasa de verdaderos negativos (TN), la proporción de negativos que el modelo predice correctamente.

$$\text{Sensibilidad} \quad \frac{TP}{TP+FN}$$

$$\text{Especificidad} \quad \frac{TN}{TN+FP}$$

- Sensibilidad y especificidad están acotados entre 0 y 1, donde valores altos implican mejor performance.

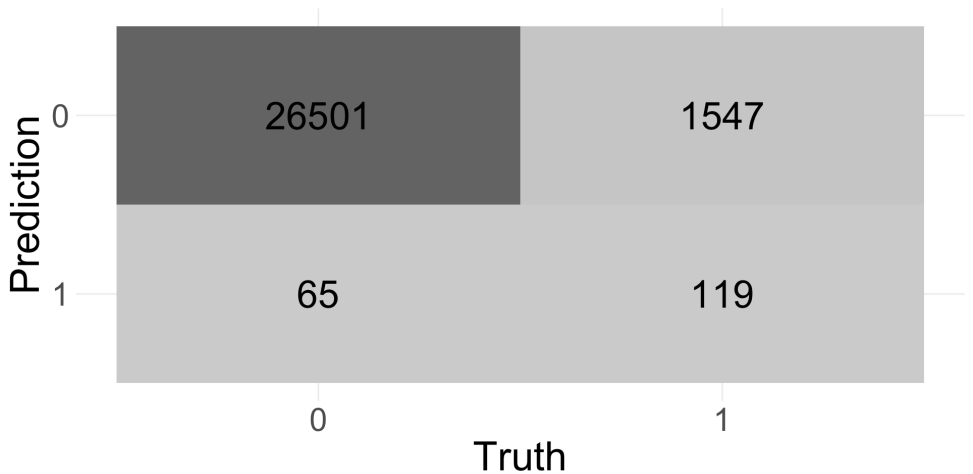
Sensibilidad y especificidad



- Sensibilidad: $\frac{\text{TP}}{\text{TP+FN}} = \frac{119}{1547+119} = 0.07$
- Especificidad: $\frac{\text{TN}}{\text{TN+FP}} = \frac{26501}{26501+65} = 0.9$

Sensibilidad y especificidad

Sensibilidad: $\frac{TP}{TP+FN}$ Especificidad: $\frac{TN}{TN+FP}$




```
# A tibble: 3 × 2
  .metric  .estimate
  <chr>        <dbl>
1 accuracy    0.943
2 sens        0.0714
3 spec        0.998
```

- 94% correctamente clasificados (precisión)
- Solo 7% de los 1s correctamente clasificados (sensibilidad)
- Casi todos los 0s son correctamente clasificados, 99.8% (especificidad)

Sensibilidad y Especificidad

- En este caso el modelo hace un buen trabajo prediciendo los ceros
- Para este problema concreto el objetivo es evitar que los estudiantes abandonen, por lo que queremos una alta sensibilidad, predecir los estudiantes que abandonan con alta precisión.
- Para nuestro objetivo el modelo no hace un buen trabajo y debemos hacer algunos ajustes.
- Si ajustamos el modelo para incrementar la sensibilidad es probable que caiga también la especificidad.

Sensibilidad y Especificidad

A menudo tenemos interés en tener una medida que refleje la tasas de falsos positivos y falsos negativos FP y FN, Podemos calcular el Youden's J Index (Youden 1950), que es

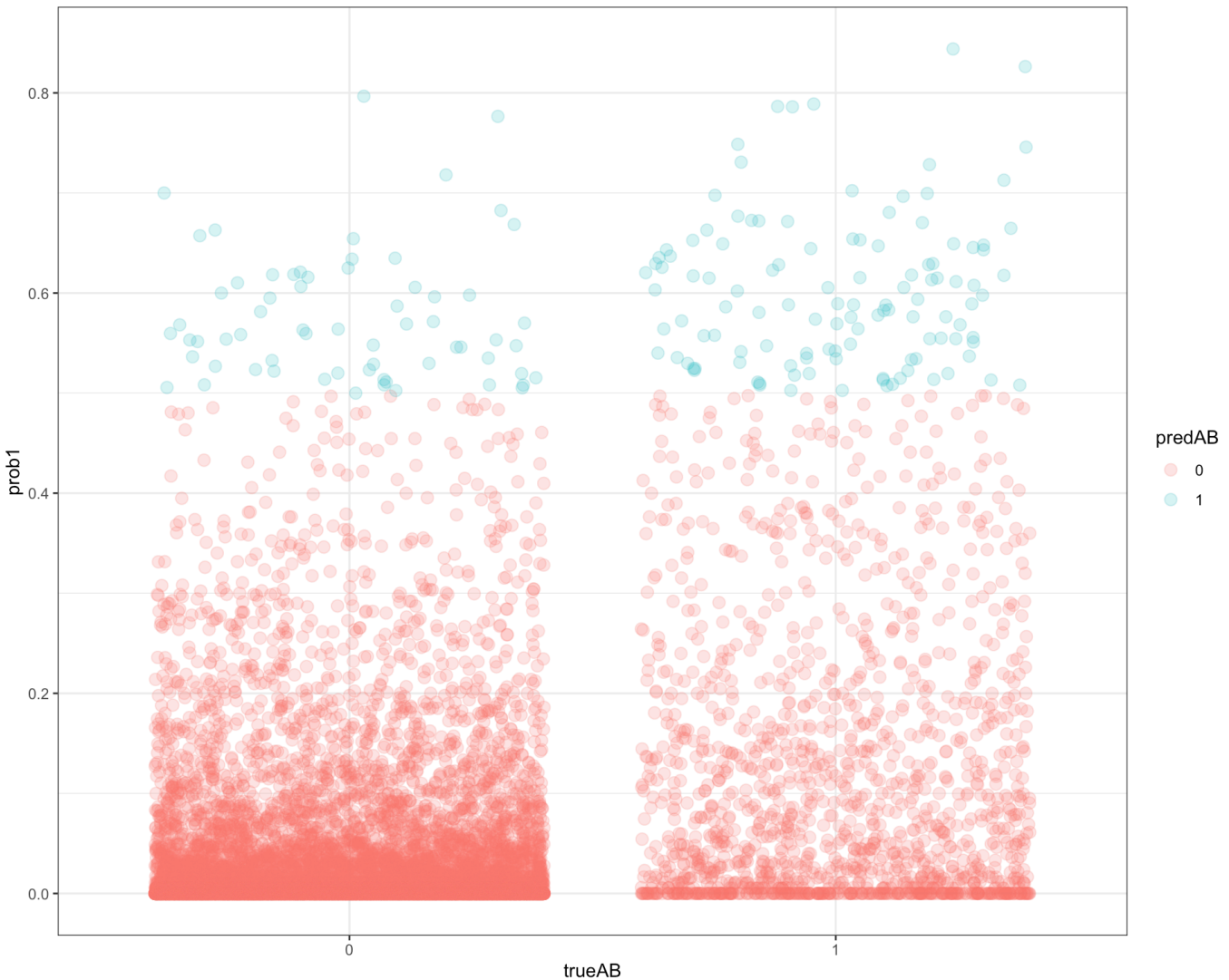
$$\text{J} = \text{Sensibilidad} + \text{Especificidad} - 1$$

- $\text{J} = 0.0714 + 0.99 - 1 = 0.061$
- Mide la proporción de muestras correctamente predichas para ambas clases. En algunos contextos es una medida apropiada para resumir la magnitud de ambos tipos de errores.
- El método más común para combinar especificidad y sensibilidad en un solo valor es el receiver operating characteristic (ROC).

Visualizando la performance del modelo

- Durante el proceso de clasificación los algoritmos estiman la probabilidad de que cada observación pertenezca a una clase particular. Conocida como propensiones y ellas se calculan teniendo en cuenta un valor de corte.
- En general para un problema de dos clases el punto de corte es 0.5.
- Es posible usar otros puntos de corte lo que afectará el cálculo de la sensibilidad y la especificidad.
- Entender como cambia la sensibilidad y la especificidad de un clasificador cuando cambia el punto de corte nos da un mejor entendimiento de la performance del modelo.
- Este balance se resume en la **Curva ROC**

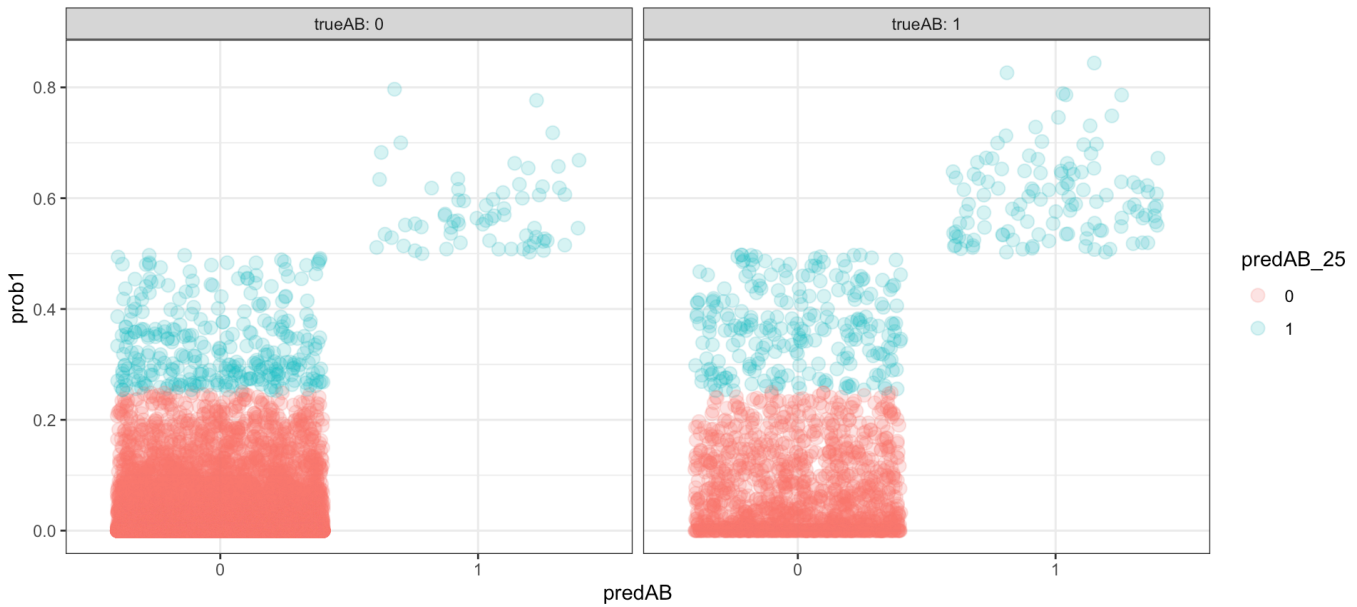
Punto de corte




```
# A tibble: 3 × 2
  .metric
.estimate
  <chr>
<dbl>
1 accuracy
0.943
2 sens
0.071
3 spec
0.998
```

Punto de corte

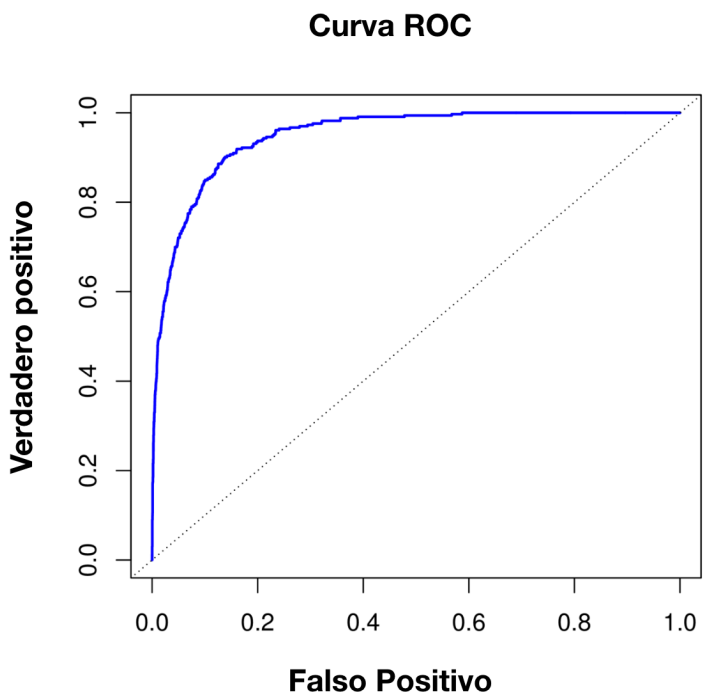
Si la probabilidad de abandono es mayor a .25 predigo abandono




```
# A tibble: 3 × 2
  .metric  .estimate
  <chr>        <dbl>
1 accuracy     0.940
2 sens         0.986
3 spec         0.217
```

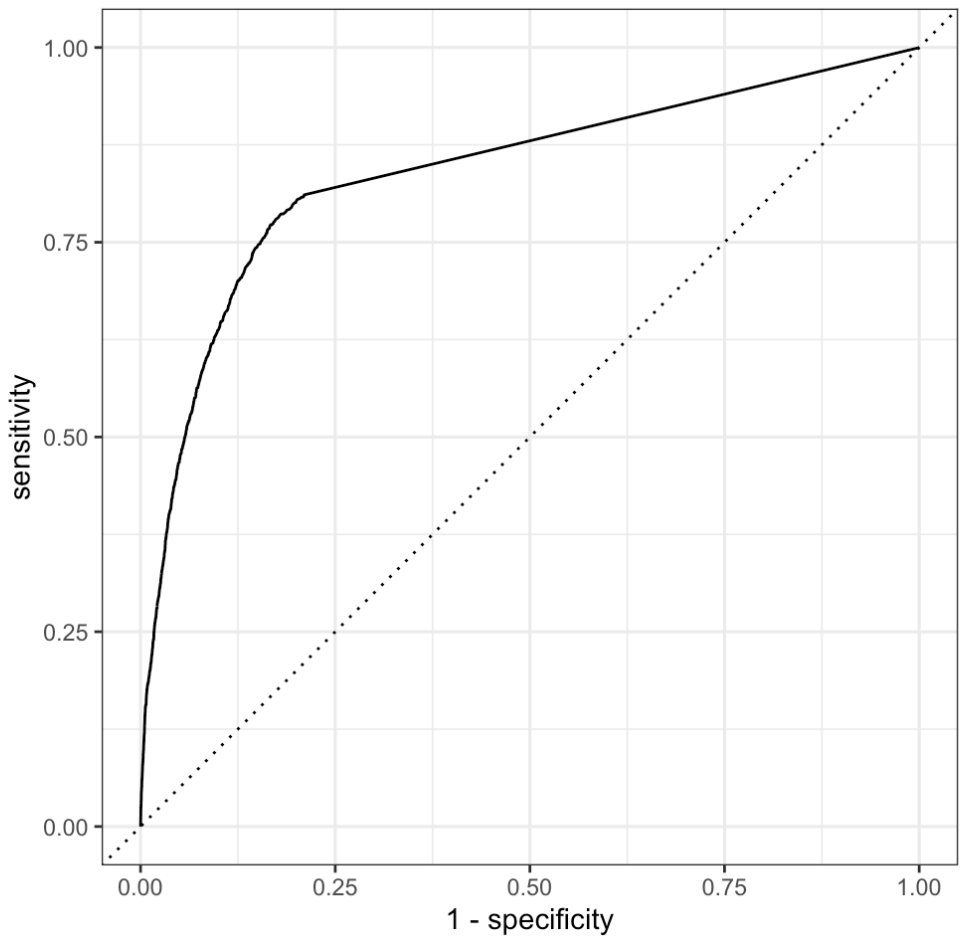
Curva ROC y AUC

Es un método común para comparar métodos de clasificación



- Eje vertical: Sensibilidad(verdaderos positivos)
- Eje horizontal: 1-Especificidad (1- verdaderos negativos)
- AUC: área bajo de la curva ROC

Curva ROC


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity

– >

– >

PPV y NPV

- Un aspecto a menudo pasado por alto de la sensibilidad y la especificidad es que son medidas condicionales. La sensibilidad es la tasa de precisión solo para la población que abandona (y la especificidad para los que no abandonan).
- Muchas veces queremos responder preguntas que no se responden con medidas condicionales.
- Queremos saber cuales son las chances que un estudiante abandone.
- Esto se responde considerando la **especicficidad** la **sensibilidad** y la **prevalencia** del evento en la población.

PPV y NPV

- Intuitivamente, si el evento es raro, esto debería reflejarse en la respuesta.
- Teniendo en cuenta la prevalencia, el análogo de la sensibilidad es el valor predictivo positivo (PPV), y el análogo de la especificidad es el valor predictivo negativo (NPV). Estos valores hacen evaluaciones incondicionales de los datos.
- El valor predictivo positivo responde a la pregunta: ¿cuál es la probabilidad de que esta muestra sea un evento?

PPV y NPV

$$\text{PPV} = \frac{\text{Sensibilidad} * \text{Prevalencia}}{(\text{Sensibilidad} * \text{Prevalencia}) + (1 - \text{Especificidad}) * (1 - \text{Prevalen}}$$

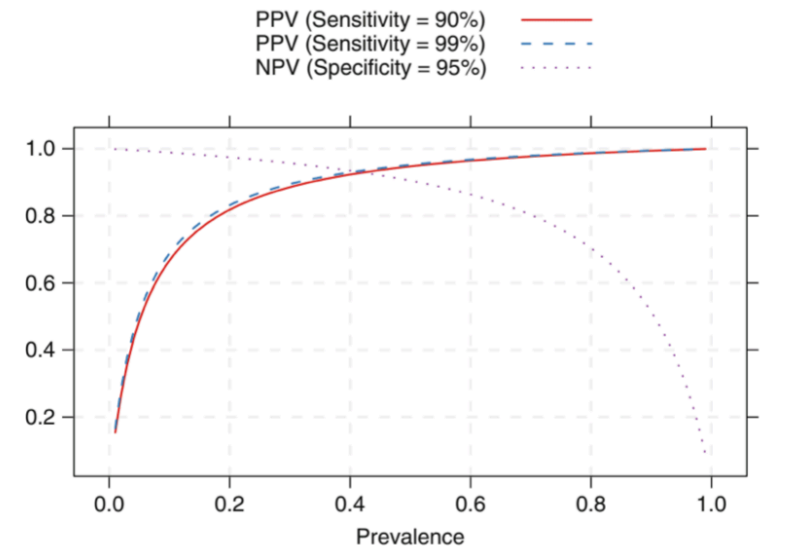
$$\text{NPV} = \frac{\text{Especificidad} * (1 - \text{Prevalencia})}{(\text{Prevalencia} * (1 - \text{Sensibilidad})) + (\text{Especificidad}) * (1 - \text{Preval}}$$

Claramente los valores predichos son combinaciones no triviales de la performance y la tasa del evento

En el ejemplo de abandono, Gracias Lucas!

PVP: 0.305 NPV: 0.944

PPV y NPV



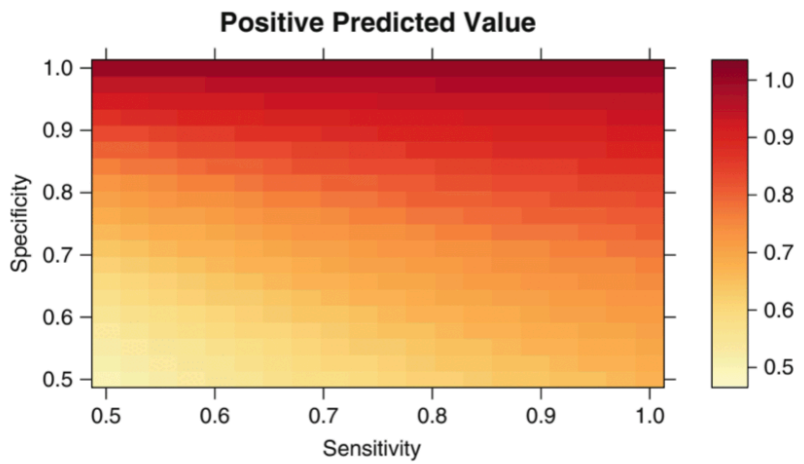


Fig. 11.5: *Top*: The effect of prevalence on the positive and negative predictive values. The PPV was computed using a specificity of 95 % and two values of sensitivity. The NPV was computed with 90 % sensitivity and 95 % specificity. *Bottom*: For a fixed prevalence of 50 %, positive predictive values are shown as a function of sensitivity and specificity

- Valores predictivos negativos grandes (NVP) pueden ser logrados cuando la prevalencia es pequeñae.
- Sin embargo cuando el evento se incrementa (prevalencia) los valores negativos predichos se vuelven pequeños.

Ejemplo abandono

- Incluso cuando el método presenta una precisión global de 94 % cuando miramos la performance contra los ejemplos positivos o negativos la perspectiva cambia.
- Es importante notar que hay varias formas de evaluar la performance de un modelo. La clave es evaluar el modelo basado en su utilidad. Es decir que la medida de performance usada para evaluar debería tomar en cuenta el propósito u objetivo del mismo.
- Veremos otras medidas de la performance del modelo que van más alla de las métricas básicas de performance predictiva.

Kappa

- En el ejemplo de abandono, si usamos una aproximación que clasifique solo usando una muestra estratificada simple, la distribución de las clases predichas es simlar a la distribución de la clase en el training entonces cuanto más desbalanceados los datos es más probable que ese predictor tenga más precisión ya que predice la mayoría de las veces la clase más probable.
- En los casos que tenemos desbalance hay que tomar en cuenta el mismo, una forma de corregir las predicciones correctas sólo por tener alta probabilidad se puede usar el Kappa de Cohen como medida de performance.
- Kappa se puede pensar como un ajuste a la precisión teniendo en cuenta que el modelo prediga correctamente sólo por que la clase tenga alta probabilidad de ser observada en la muestra.

Kappa

- Para ello debemos calcular la probablidad esperada de acierto (pe) entre el valor predicho y el real bajo el supuesto que las predicciones fueron realizadas aleatoriamente.
- Luego usamos esta medida para ajustar la precisión (pa) del modelo.

Kappa se calcula: $\varkappa = \frac{\text{pa}-\text{pe}}{1-\text{pe}}$

Donde $\text{pa} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} = \frac{119+26501}{119+26501+65+1547} = 0.942$

Debemos calcular la probablidad esperada de acierto (pe). La probabilidad que los valores predichos y esperados coincidan.

Kappa

La conjunta de que el valor predicho y observad sean abandono.

- La probabilidad de que se prediga abandono es $\frac{65+119}{28232} = 0.0065$
- La probabilidad que un estudiante abandone es $\frac{1547+119}{28232} = 0.059$
- Probabilidad de que predichos y observados abandonen $0.0065 * 0.059 = 0.00038$.
- De similar manera se construye para predichos y observados que no abandonen. $(0.99 * 0.94 = 0.93)$

Kappa

- Predicho y real de abandono son mutuamente excluyente del predicho y el observado de no abandono entonces la probabilidad de acuerdo o de abandono o no abandono es la suma de ambas probabilidades.
- Esto significa que $pe = 0.00038 + 0.93 = 0.90$.

Entonces: $\varkappa = \frac{0.942 - 0.90}{1 - 0.90} = 0.42$

- Significa que la precisión predicha del modelo ajustada por predicciones correctas solo por chance es 0.42%. $\varkappa$ va de -1 a 1
- Si es 0.5 indica performance moderada a muy buena y menor que 0.5 indica justa a muy pobre.

Calibration Plot

Una forma de evaluar la calidad de las probabilidades de clase es utilizando un gráfico de calibración. Para un conjunto de datos dado, este gráfico muestra alguna medida de la probabilidad observada de un evento frente a la probabilidad de clase predicha.

Calibration Plot

- Una forma de crear esta visualización es evaluar un conjunto de muestras con resultados conocidos (preferiblemente un conjunto de prueba) utilizando un modelo de clasificación.
- El siguiente paso es agrupar los datos en intervalos basados en sus probabilidades de cada clase. Por ejemplo, un conjunto de intervalos podría ser $[0, 10\%], (10\%, 20\%], \ldots, (90\%, 100\%]$.
- Para cada intervalo, determina la tasa de eventos observada. Supón que 50 muestras cayeron en el intervalo de probabilidades de clase menores al 10% y hubo un solo evento.
- El punto medio del intervalo es 5% y la tasa de eventos observada sería 2%.
- El gráfico de calibración mostraría el punto medio del intervalo en el eje x y la tasa de eventos observada en el eje y. Si los puntos caen a lo largo de una línea de 45°, el modelo ha producido probabilidades bien calibradas.

Calibration Plot

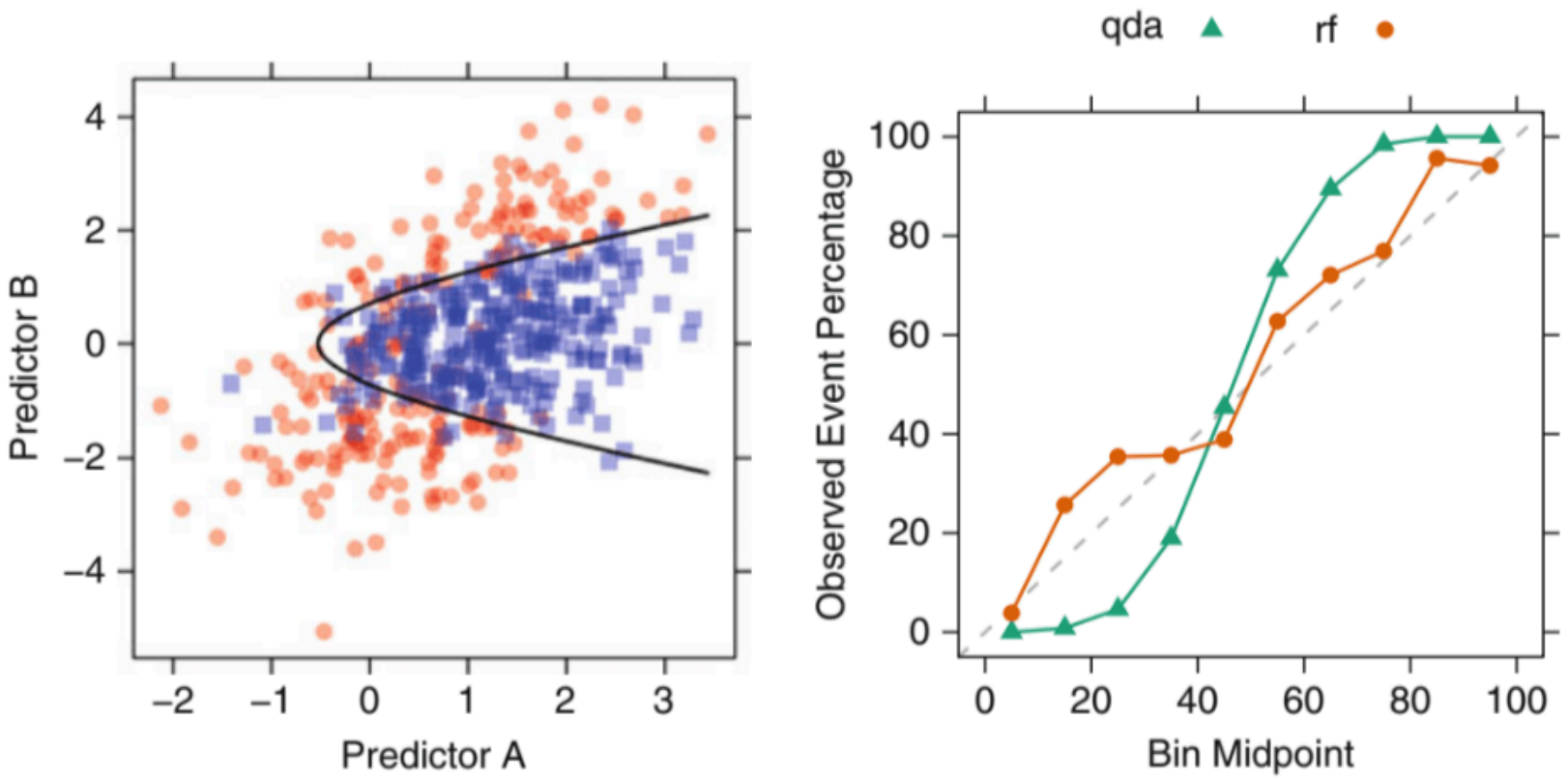


Fig. 11.1: *Left*: A simulated two-class data set with two predictors. The *solid black line* denotes the 50 % probability contour. *Right*: A calibration plot of the test set probabilities for random forest and quadratic discriminant analysis models