Selección y Validación de Modelos

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Junio de 2020

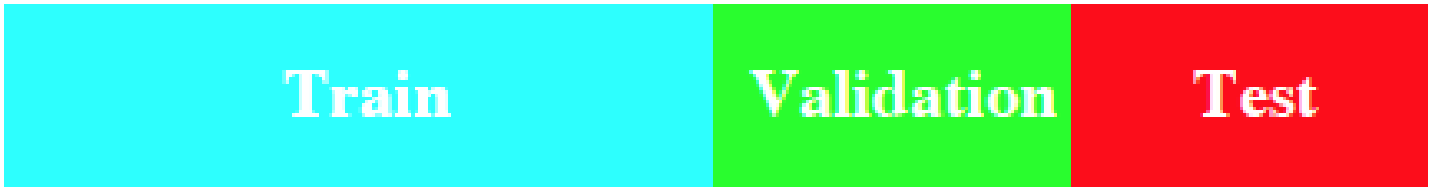
Contenido

Objetivos

- Elegir parámetros (meta parámetros).
- Estimar el error de prueba.

Estimación Error

- En ambientes ricos en datos podemos seleccionar modelos (metaparámetros) y estimar el riesgo.



Estimación Error

- Si bien el mayor interés es la estimación del riesgo (condicional a una muestra), esto se puede hacer solo cuando hay muchos datos.
- En general muy lo máximo que se puede hacer es estimar el riesgo esperado.

Contenido

Validación Cruzada

- Cuando no se tiene tanta información se puede hacer validación cruzada.
- Esto permite estimar el riesgo esperado y seleccionar modelos.
- Validación cruzada de K muestras:
  1. Dividir en K muestras aleatorias la muestra original de tamaño N. Dada la muestra k se entrena el modelo sin los datos de esta muestra y se estima el error en esa muestra. El promedio de los errores es la estimación del riesgo esperado.
  2. Cuando $K = N$ se conoce como *leave out one cross validation*. En este caso el modelo estimado puede tener una varianza alta pero el sesgo en la estimación del error esperado es bajo.

Validación Cruzada: Validación Cruzada: K vrs. 1

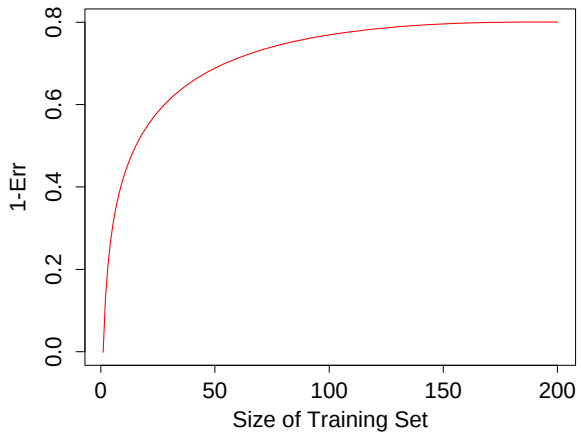


FIGURE 7.8. *Hypothetical learning curve for a classifier on a given task: a plot of* $1-\text{Err}$ *versus the size of the training set* N*. With a dataset of* 200 *observations,* 5*-fold cross-validation would use training sets of size* 160*, which would behave much like the full set. However, with a dataset of* 50 *observations fivefold cross–validation would use training sets of size* 40*, and this would result in a considerable overestimate of prediction error.*

- El número de divisiones óptimo depende del número de datos. Un K alto con pocos datos sobre estima el error de prueba (varianza alta). Un K bajo subestima el error (sesgo alto).

Validación Cruzada: Uso adecuado

- Considere un problema con mucho predictores.
- Haga una selección de variables con base en la correlación (univariada) de las variables con la variable objetivo.
- Usando este subconjunto construya un clasificador.
- Use validación cruzada

Validación Cruzada: Uso adecuado

- Considere el siguiente ejemplo: una muestra de tamaño 50 con el mismo número de dos clases. 5000 predictores independientes de las clases. El mejor modelo tiene un error de clasificación del 50 %.
- Al usar la estrategia anterior eligiendo los 100 predictores que más correlación tienen con la variable objetivo y después usando un clasificador de un vecino más cercano se hace validación cruzada con $K = 5$.
- Si se simula 50 veces este procedimiento el error de validación cruzada es 3 %.

Validación Cruzada: Uso adecuado

- El problema es que la selección devariables se hace con todos los datos incluso aquellos que se van a utilizar para hacer la validación. Esto es un problema clasico de fuga de información (i.e, information leakeage).
- En la muestra de validación la correlación es muy alta.

Validación Cruzada: Uso adecuado

- La forma correcta es: hacer la selección de variables como parte de la validación cruzada.
- En general cuando hay varias etapas en la metodología en cada remuestreo se debe aplicar la metodología.
- La única excepción es cuando se ha usado algún método para filtrar que no sea supervisado (no dependa de la variable objetivo). Por ejemplo, elegir aquellas variables con mayor varianza.

Validación Cruzada: Uso adecuado

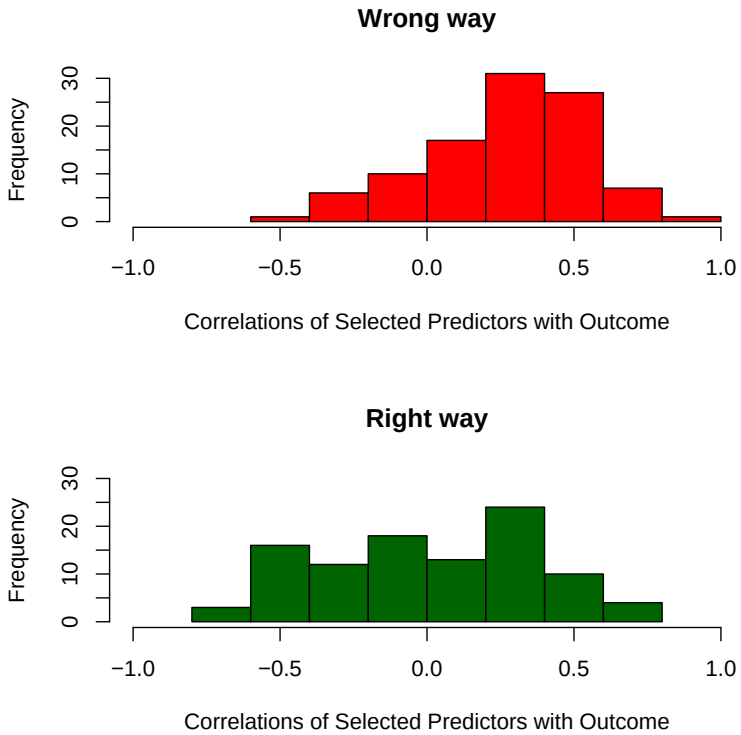


FIGURE 7.10. *Cross-validation the wrong and right way: histograms shows the correlation of class labels, in* 10 *randomly chosen samples, with the* 100 *predictors chosen using the incorrect (upper red) and correct (lower green) versions of cross-validation.*

Contenido

Bootstrap

- En general la metodología de boostrap se utiliza para deducir las propiedades estadísticas de funciones de una muestra finitas (estimadores).
- Aquí vamos a explicar como usarlo para estimar el riesgo esperado.

Bootstrap: Estadísticos muestras finitas

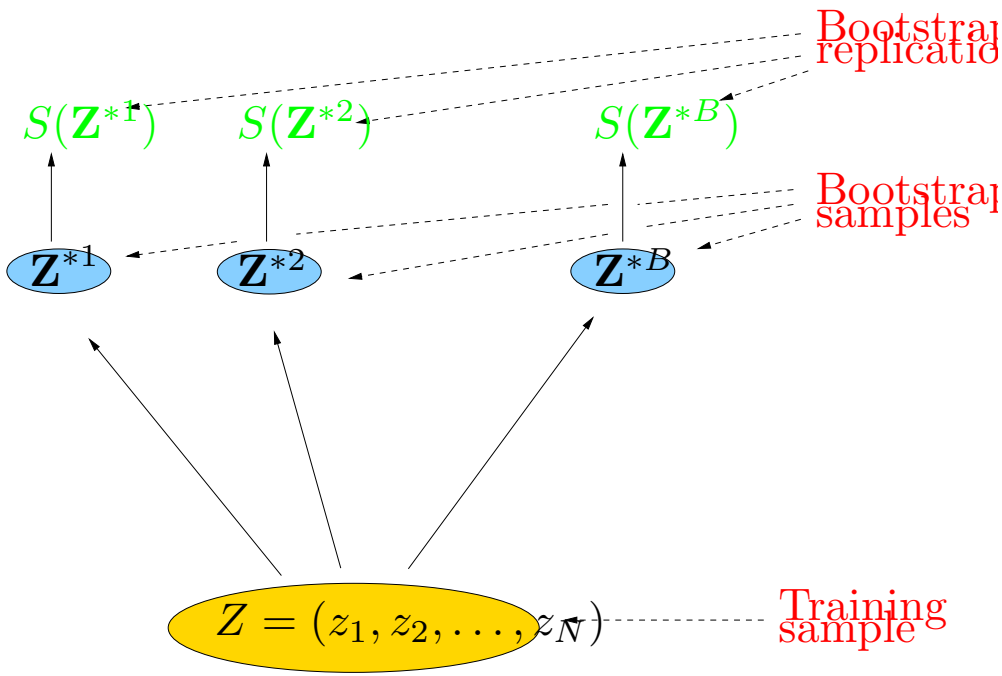


FIGURE 7.12. *Schematic of the bootstrap process. We wish to assess the statistical accuracy of a quantity $S(\mathbf{Z})$ computed from our dataset. B training sets $\mathbf{Z}^{*b}$, $b = 1, \ldots, B$ each of size N are drawn with replacement from the original dataset. The quantity of interest $S(\mathbf{Z})$ is computed from each bootstrap training set, and the values $S(\mathbf{Z}^{*1}), \ldots, S(\mathbf{Z}^{*B})$ are used to assess the statistical accuracy of $S(\mathbf{Z})$.*

Contenido

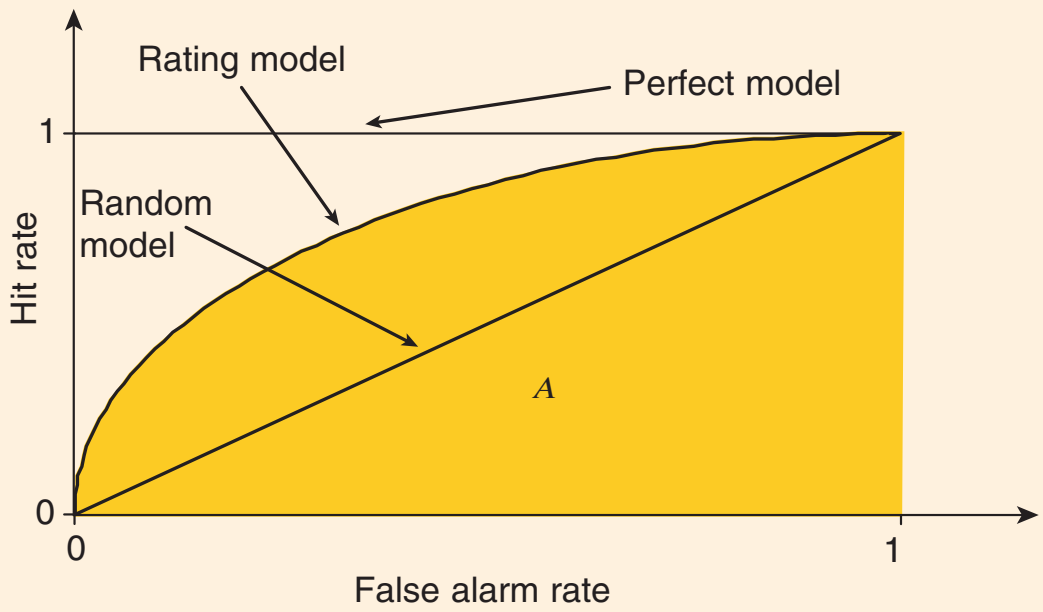
Validación de Modelos

- Modelos de regresión: AIC, R^2, MAPE, etc.
- Modelo de clasificación: Curva ROC y curva de calibración

Curva ROC

- La curva ROC y el área bajo la curva es uno de los métodos más importantes de validación para problemas de clasificación.

3. Receiver operating characteristic curves



Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la dsitribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distacia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene lo que algunos libros llaman la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vrs. (1-Especificidad (x)).
- La distacia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.

Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la dsitribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distacia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene lo que algunos libros llaman la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vrs. (1-Especificidad (x)).
- La distacia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.

Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la dsitribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distacia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene lo que algunos libros llaman la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vrs. (1-Especificidad (x)).
- La distacia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.

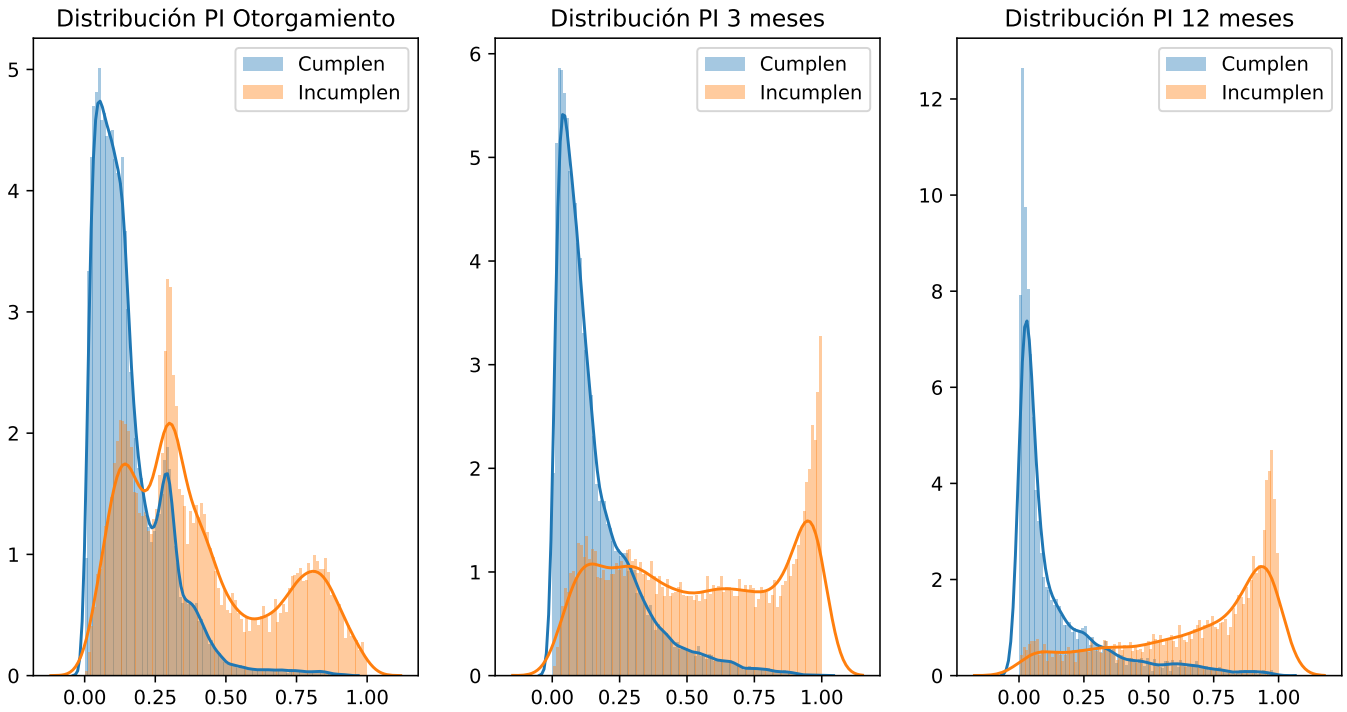
Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la dsitribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distacia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene lo que algunos libros llaman la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vrs. (1-Especificidad (x)).
- La distacia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.

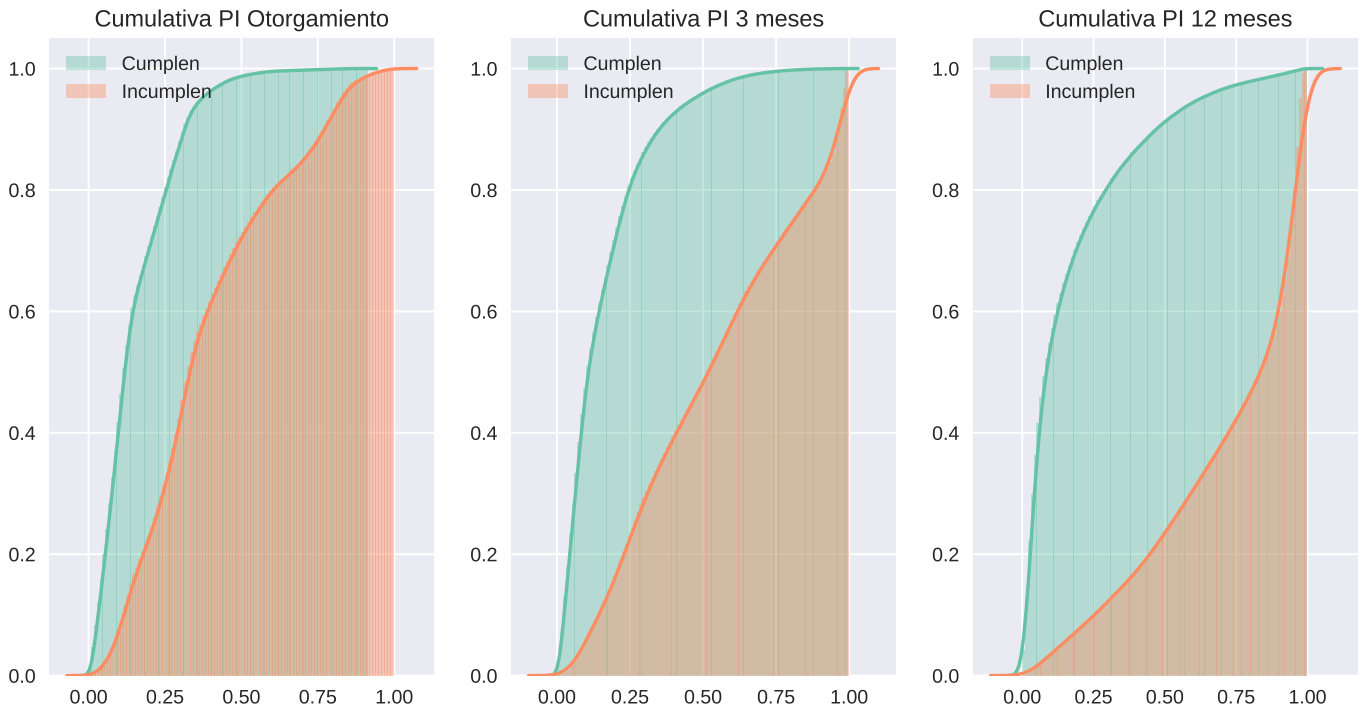
Curva ROC

- Por simplicidad definamos el score de un individuo como la probabilidad de incumplimiento.
- Considere una gráfica de la dsitribución acumulada del score de los buenos y los malos. El score que arroje la distancia máxima entre estas distribuciones es la distacia de Kolmogorov y Smirnov.
- Si representamos estas dos gráficas en un mismo gráfico se obtiene lo que algunos libros llaman la curva ROC: en el eje y la función de distribución de los buenos y en el eje x la función de distribución de los malos: Sensibilidad (y) vrs. (1-Especificidad (x)).
- La distacia KS corresponde al score donde la distancia horizontal entre la curva ROC y la diagonal es máxima (pendiente 1).
- El coeficiente de Gini es el doble del área entre la diagonal y la curva ROC.

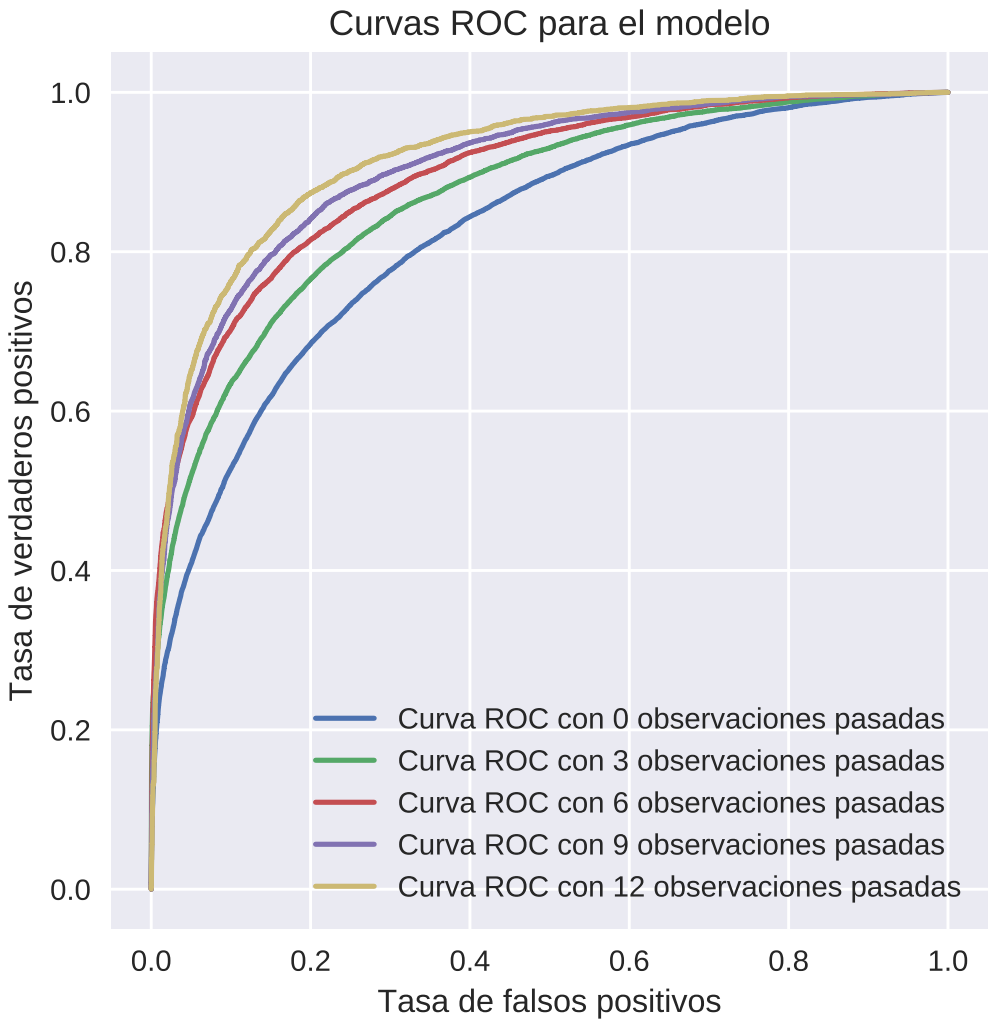
Curva ROC


Distribución PI Otorgamiento
Cumplen
Incumplen
0
1
2
3
4
5
0.00
0.25
0.50
0.75
1.00
Distribución PI 3 meses
Cumplen
Incumplen
0
1
2
3
4
5
6
0.00
0.25
0.50
0.75
1.00
Distribución PI 12 meses
Cumplen
Incumplen
0
2
4
6
8
10
12
0.00
0.25
0.50
0.75
1.00

Curva ROC


Cumulativa PI Otorgamiento
Cumplen
Incumplen
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
Cumulativa PI 3 meses
Cumplen
Incumplen
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
Cumulativa PI 12 meses
Cumplen
Incumplen
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00

Curva ROC


Curvas ROC para el modelo
Tasa de verdaderos positivos
Tasa de falsos positivos
0.0
0.2
0.4
0.6
0.8
1.0
Curva ROC con 0 observaciones pasadas
Curva ROC con 3 observaciones pasadas
Curva ROC con 6 observaciones pasadas
Curva ROC con 9 observaciones pasadas
Curva ROC con 12 observaciones pasadas

Curva ROC: Explicación

- Sea M el total de unos, B el total de ceros y $T = M + B$.
- La tasa de verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M.
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B.
- Entre más cerca este la curva del eje vertical y el eje horizontal superior, mejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.

Curva ROC: Explicación

- Sea M el total de unos, B el total de ceros y $T = M + B$.
- La tasa de verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M.
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B.
- Entre más cerca este la curva del eje vertical y el eje horizontal superior, mejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.

Curva ROC: Explicación

- Sea M el total de unos, B el total de ceros y $T = M + B$.
- La tasa de verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M.
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B.
- Entre más cerca este la curva del eje vertical y el eje horizontal superior, mejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.

Curva ROC: Explicación

- Sea M el total de unos, B el total de ceros y $T = M + B$.
- La tasa de verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M.
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B.
- Entre más cerca este la curva del eje vertical y el eje horizontal superior, mejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.

Matriz de confusión

Obs\Pred	m	b
M	vp	fn
B	fp	vn

Matriz de confusión

- Tasa de verdaderos positivos: $TVP = \frac{vp}{M}$. También se conoce como *recall*, *hit rate*, *sensitivity*.
- Tasa de falsos positivos: $TFP = \frac{fp}{B}$. También se conoce como *false alarm rate* (1-*specificity*).

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x. El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como 1 si $f(x) \geq 1$.
- Entonces:

$$TVP = \frac{(\frac{M}{T} \times (1 - s)T)}{M} = 1 - s$$

$$TFP = \frac{(\frac{B}{T} \times (1 - s)T)}{B} = 1 - s$$

- Luego la curva ROC es la diagonal.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x. El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como 1 si $f(x) \geq 1$.
- Entonces:

$$TVP = \frac{(\frac{M}{T} \times (1 - s)T)}{M} = 1 - s$$

$$TFP = \frac{(\frac{B}{T} \times (1 - s)T)}{B} = 1 - s$$

- Luego la curva ROC es la diagonal.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x. El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como 1 si $f(x) \geq 1$.
- Entonces:

$$TVP = \frac{(\frac{M}{T} \times (1 - s)T)}{M} = 1 - s$$

$$TFP = \frac{(\frac{B}{T} \times (1 - s)T)}{B} = 1 - s$$

- Luego la curva ROC es la diagonal.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x. El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como 1 si $f(x) \geq 1$.
- Entonces:

$$TVP = \frac{(\frac{M}{T} \times (1 - s)T)}{M} = 1 - s$$

$$TFP = \frac{(\frac{B}{T} \times (1 - s)T)}{B} = 1 - s$$

- Luego la curva ROC es la diagonal.

Curva ROC: Explicación

- El modelo aleatorio tiene como curva ROC la diagonal de 45 grados.
- Para ver esto fijemos un score s y supongamos que cuando tenemos un ejemplo con características x. El modelo aleatorio le asigna un score $f(x) \in [0, 1]$ que se distribuye uniforme y se marca como 1 si $f(x) \geq 1$.
- Entonces:

$$TVP = \frac{(\frac{M}{T} \times (1 - s)T)}{M} = 1 - s$$

$$TFP = \frac{(\frac{B}{T} \times (1 - s)T)}{B} = 1 - s$$

- Luego la curva ROC es la diagonal.

Curva de calibración

- Mide el error entre las frecuencias pronosticadas de un evento y las frecuencias observadas de ocurrencia.
- En las aplicaciones de aprendizaje de máquinas, a medicina es estándar y se usa una prueba χ^2 para determinar la significacia estadística de la diferencia.
- Es manipulable.

Curva de calibración

- Mide el error entre las frecuencias pronosticadas de un evento y las frecuencias observadas de ocurrencia.
- En las aplicaciones de aprendizaje de máquinas, a medicina es estándar y se usa una prueba χ^2 para determinar la significacia estadística de la diferencia.
- Es manipulable.

Curva de calibración

- Mide el error entre las frecuencias pronosticadas de un evento y las frecuencias observadas de ocurrencia.
- En las aplicaciones de aprendizaje de máquinas, a medicina es estándar y se usa una prueba χ^2 para determinar la significacia estadística de la diferencia.
- Es manipulable.