FACULTAD DE INGENIERÍA Y ARQUITECTURA
CARRERA DE INGENIERÍA DE SISTEMAS

ANÁLISIS DE REGRESIÓN LOGÍSTICA

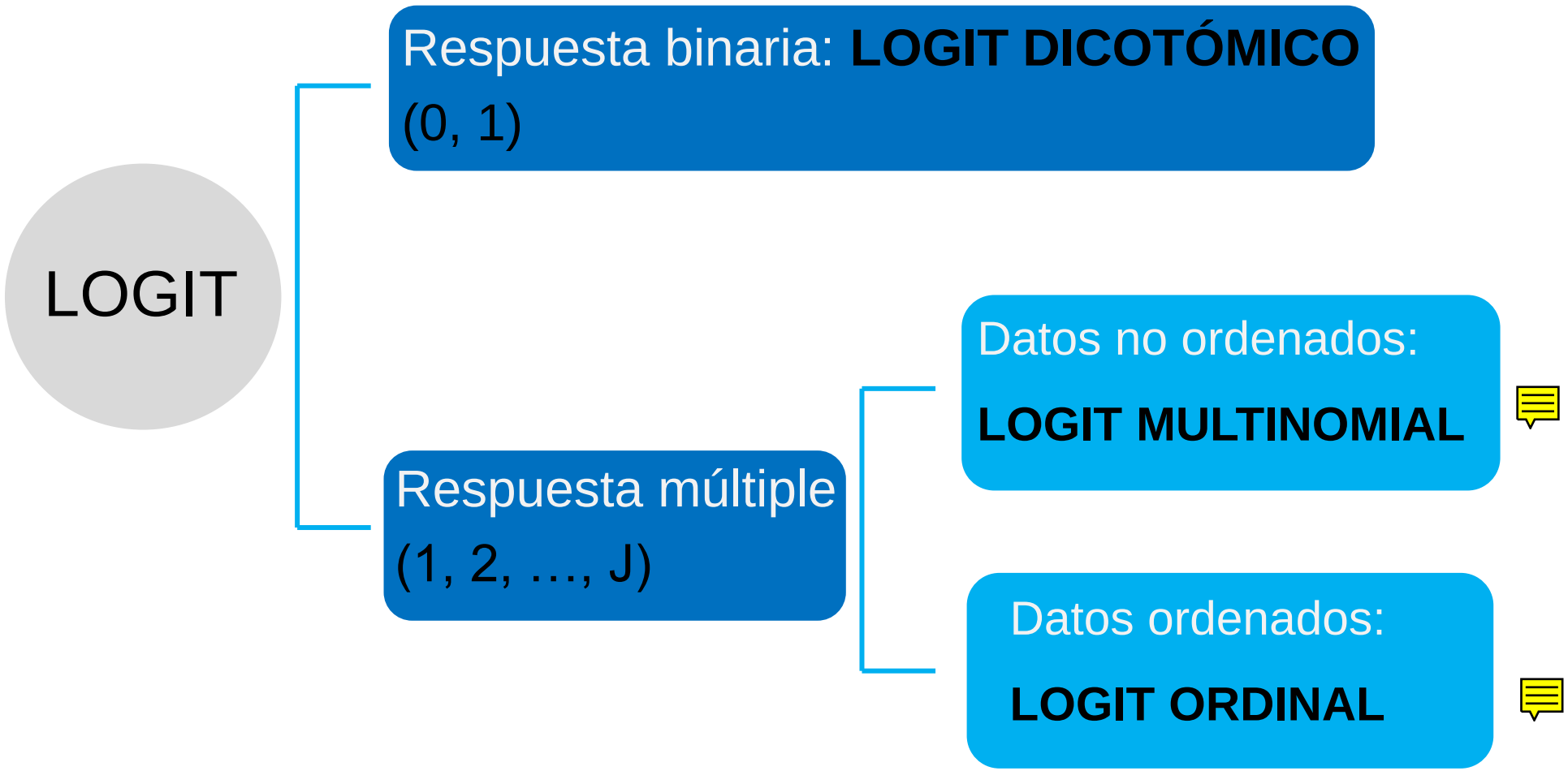
ANALÍTICA PREDICTIVA
PROFESORES DEL CURSO:
MSC. ALDO MEZA RODRÍGUEZ

Modelo de regresión Logística

Logística
Regresión

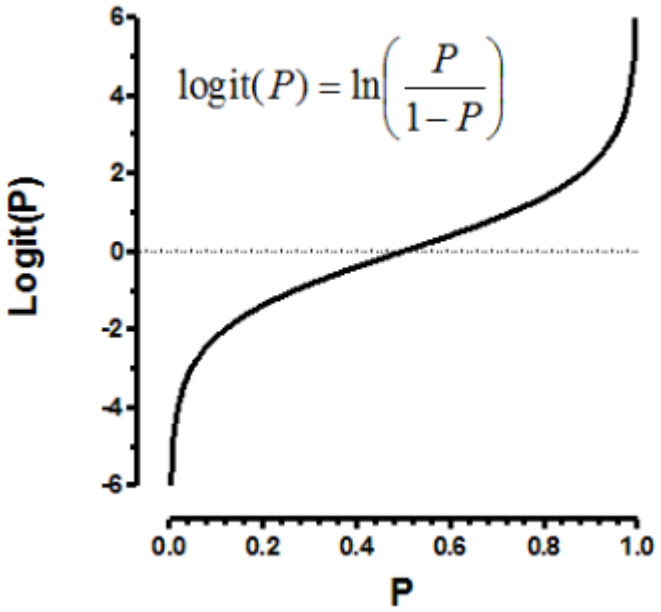
- Son modelos de regresión donde la variable dependiente es de carácter cualitativo y las independientes cuantitativas o cualitativas.
- En su forma básica utiliza una función logística para modelar una variable dependiente binaria.
- Requieren realizar menos supuestos, lo que permite, obtener unos resultados más robustos.

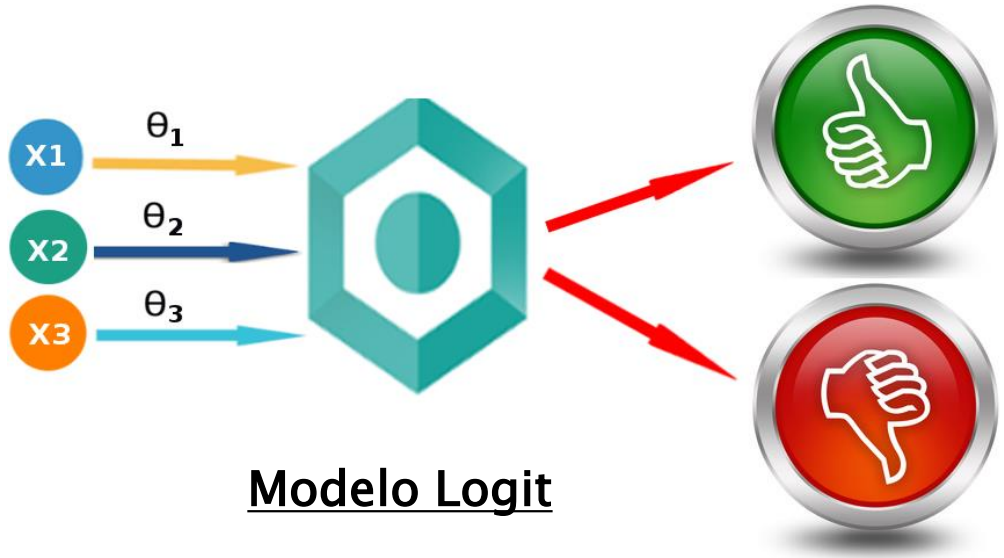
TIPOLOGÍA DE MODELOS LOGIT



LA REGRESIÓN LOGÍSTICA

La regresión logística es un modelo estadístico que en su forma básica utiliza una función logística para modelar una variable dependiente binaria





Modelo Logit

$$Y_i = \frac{1}{1+e^{-\alpha-\beta_k X_{ki}}} + u_i = \frac{e^{\alpha+\beta_k X_{ki}}}{1+e^{\alpha+\beta_k X_{ki}}} + u_i$$

El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM) que usa como función de enlace la función logit.

Regresión logística binaria

- Modela cómo influye en la probabilidad de aparición de un suceso, la presencia o no de diversos factores.
- El modelo logístico se puede linealizar mediante logits de las probabilidades binomiales desconocidas (los logaritmos de la razón de momios)
- Los parámetros desconocidos β son usualmente estimados a través del método de máxima verosimilitud
- El modelo logístico establece una relación entre la probabilidad de que ocurra el suceso, dado que el individuo presenta los valores X_{ki}
- Los coeficientes de regresión logística pueden utilizarse para estimar la razón de las ventajas (odds Ratio OR) de cada variable independiente del modelo.

$$\pi(Y_i = 1) = \frac{e^{\alpha+\beta_k X_{ki}}}{1+e^{\alpha+\beta_k X_{ki}}}$$

$$\ln[\frac{p}{1\text{-}p}] = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

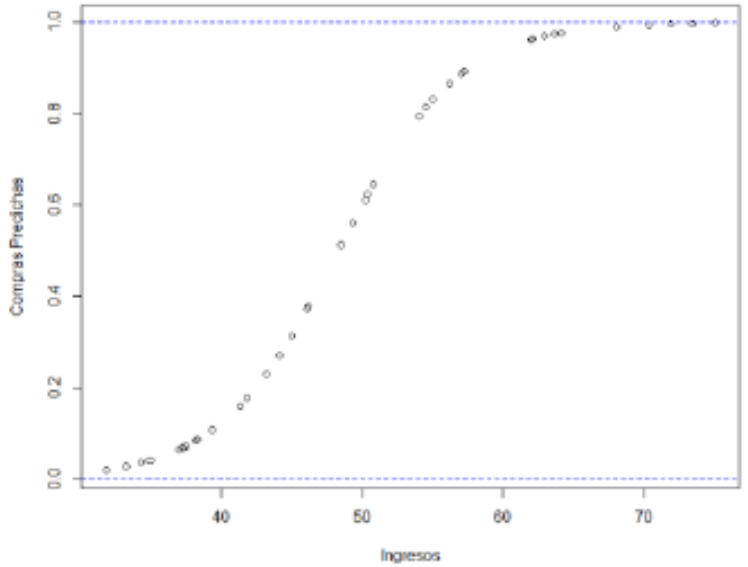
$$odds = \frac{p}{1-p}$$

$$OR = \exp(\beta)$$

Modelos Logit

```
(Intercept)     ingresos
 -11.449161     0.237138
```

$$\hat{p}_i = \frac{e^{-11.449+0.2371(ingresos)}}{1+e^{-11.449+0.2371(ingresos)}} \qquad 1-\hat{p}_i = \frac{1}{1+e^{-11.449+0.2371(ingresos)}}$$



Etapas para construir un modelo logístico

1. **Especificación**

Definición de la variable de respuesta (variable dependiente) y las variables explicativas (independientes)

2. **Estimación**

Estimación de los parámetros y el modelo

3. **Validación**

Conjunta: Ver si en forma global el modelo es aceptable. **Individual:** Ver si al menos una variable es significativa). Analizar la bondad de ajuste

4. **Utilización**

Predicción, evaluación de métricas, tasas de predicción correcta e interpretación de parámetros

Utilización del modelo

Contraste de significación para todos los β_j

Para verificar: $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$
$H_1 : \text{Al menos un } \beta_j \neq 0$

$$RV_0 = -2\left[\ln L_0 - \ln L\right]$$

lnL: el logaritmo de la función de verosimilitud que se ha obtenido al estimar el modelo completo
LnL0: es el logaritmo de la función de verosimilitud al estimar el modelo con sólo el término independiente

Cuando n tiene al infinito el estadístico de razón de varosimilitud se aproxima a una chicuadrado.

$$RV_0 = \ln L_0 - \ln L \sim \chi^2_{(k-1)}$$

Contraste de significación para un β_j

Para probar $H_0 : \beta_j = 0$ $H_1 : \beta_j \neq 0$ se utiliza el estadístico de Wald.

Wald= $\left(\frac{\hat{\beta} - \beta}{DT(\hat{\beta})}\right)^2$ = Distrib. similar a t^2 ⇨ $\left(\frac{\hat{\beta}}{DT(\hat{\beta})}\right)^2$ = Distrib. similar a t^2 si H_0 cierta

Utilización del modelo

Prueba de bondad de ajuste de Hosmer Lemeshow

H_0 : El modelo ajustado es el adecuado (no existen diferencias entre valores observados y predichos).

$$HL = \sum_{i=1}^{10} \frac{\left[O_i - N_i \bar{\pi}_i\right]^2}{N_i \bar{\pi}_i \left(1 - \bar{\pi}_i\right)} \sim \chi_8^2 \text{ donde:}$$

O_i : es el número de unos del decil i.

$\bar{\pi}_i$: es la media de las probabilidades predichas en el decil i.

N_i : es el número de observaciones del decil i.

$$\text{Pseudo } R^2 = 1 - \frac{\log L\,(completo)}{\log L\,(reducido)}$$

Pseudo R^2 de Cox y Snell:

$$R^2_{CS} = 1 - \left(\frac{L_{constante}}{L_{modelo}}\right)^{\frac{2}{n}} = 1 - \exp\left(\frac{\Lambda_{modelo} - \Lambda_{constante}}{n}\right)$$

Pseudo R^2 de Nagelkerke:

$$R^2_N = \frac{1 - \left(\frac{L_{constante}}{L_{modelo}}\right)^{\frac{2}{n}}}{1 - \left(L_{constante}\right)^{\frac{2}{n}}} = \frac{1 - \exp\left(\frac{\Lambda_{modelo} - \Lambda_{constante}}{n}\right)}{1 - \exp\left(\frac{-\Lambda_{constante}}{n}\right)}$$

Si $\Lambda = -2log(L)$ entonces $L = \exp(-\Lambda/2)$

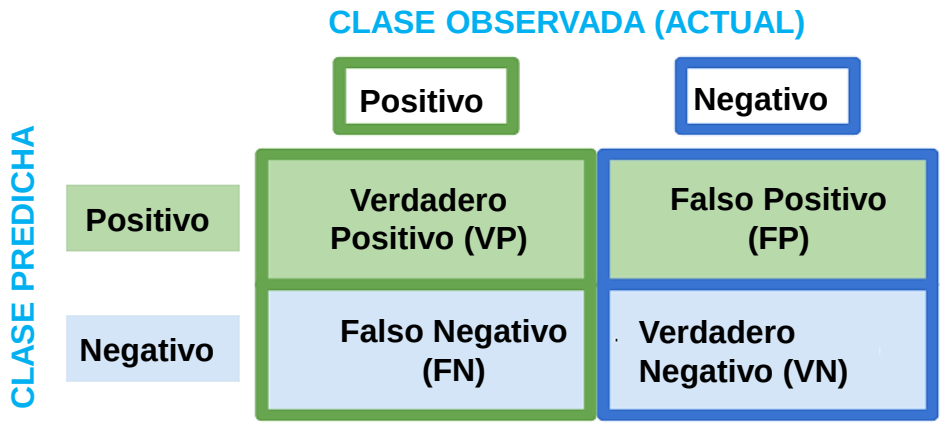
UNIVERSIDAD DE LIMA

Interpretaciones de los coeficientes

En los modelos de regresión logística, la interpretación de los coeficientes no es tan directa, debido a la naturaleza no lineal del modelo lineal.

Se debe considerar la interpretación según las posibles escalas de medida de la variable independiente.

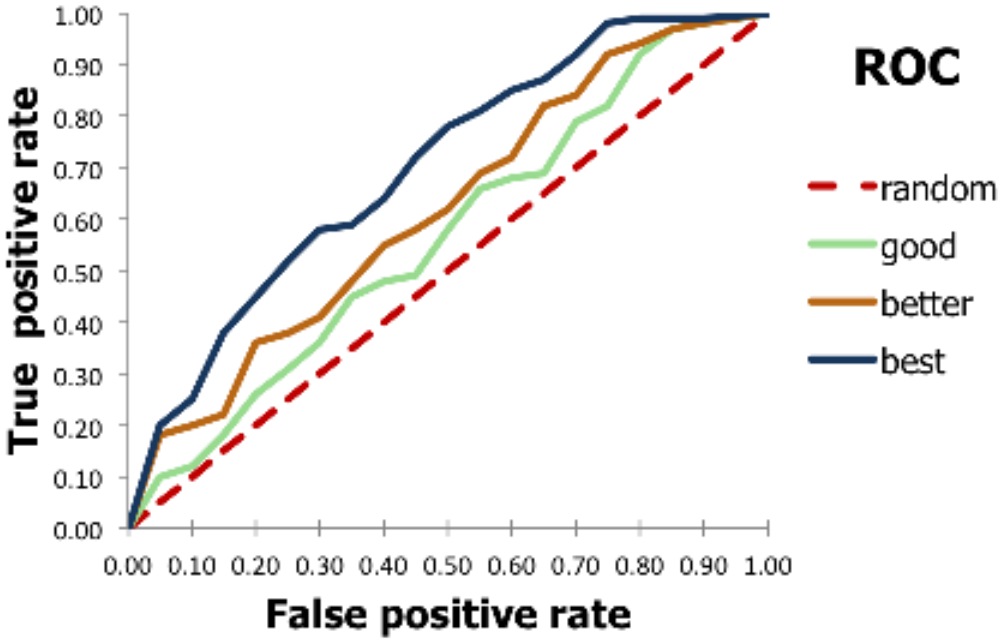
MATRIZ DE CONFUCIÓN



- **Accuracy:** Es la tasa de predicción correcta (VP + VN)/Total de casos
- **Tasa de error:** (FP + FN)/Total de casos = 1 –Accuracy
- **Sensibilidad:** Es la probabilidad de clasificar correctamente a un valor positivo. VP/(FN+VP)
- **Especificidad**: Es la probabilidad de clasificar correctamente a un valor negativo. VN/(VN+FP)

CURVAS ROC

- Se forma con la tasa verdadera positiva (sensitividad) en función de la tasa falsa positiva (1-especificidad) para diferentes puntos de corte de un modelo.
- El clasificador tiene un mejor desempeño si la curva ROC se aleja más de la diagonal principal.
- El área bajo la curva (AUC) la probabilidad de que el clasificador pronostique correctamente a dos individuos. Uno que pertenece a la categoría Y=1 y otro que pertenece a Y=0.



AUC ROC	Conclusión
0.5	No discrimina
0.6 - 0.7	Pobre
0.71 - 0.8	Aceptable
0.81 - 0.9	Excelente (bueno)
> 0.9	Muy bueno

Ejemplo regresión logística

- **Caso:** El gerente de una empresa de seguros de vida desea analizar a sus posibles clientes a través de suscripciones, es decir, desea evaluar si un determinado cliente se suscribirá o no al programa de seguro. Para ello estudia diferentes variables con la intención de identificar cuáles son determinantes para que una persona se adquiera o no el seguro.

- **Objetivo:** Predecir la suscripción o no a un seguro de vida en función a determinadas variables.

Independientes:

Edad de la persona (en años)
Sexo de la persona (1= Masculino, 0 = Femenino)
Número de hijos
Ingresos: En soles
Meses de trabajo en la empresa.
Automóvil: (1= si tiene, 0= no tiene)

Dependiente:

Seguro: 1= si, 0 = No

Pruebas global e individual

$H_0: \beta_1 = \beta_2 = \cdots = \beta_6 = 0$
$H_1:$ Al menos un $\beta_i \neq 0$

Model Likelihood Ratio Test	
LR chi2	56.67
d.f.	6
Pr(> chi2)	<0.0001

AIC: 104.08

A un nivel de significancia del 0.05 se rechaza la hipótesis planteada por lo que concluimos que al menos una de las variables predictoras influye en el éxito o fracaso de adquirir el seguro.

Prueba de Wald

H_0= ß$_j$ = 0 (la información que se perdería al eliminar X_j en el siguiente caso no es significativa)

H1= ßj ≠ 0

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.6959933	1.8895157	-3.544	0.000394	***
Sexo	0.6630229	0.5285309	1.254	0.209673	
Edad	0.1262154	0.0309603	4.077	4.57e-05	***
Hijos	0.3978139	0.2215466	1.796	0.072555	.
Ingresos	-0.0002102	0.0003399	-0.618	0.536287	
Empresa	-0.0100856	0.0040583	-2.485	0.012949	*
Automóvil	1.8735108	0.6494037	2.885	0.003914	**

Se aprecia que las variables sexo e ingresos, junto con el sexo para la categoría masculino no son significativos, por lo que no se perdería información si se eliminan del modelo

Pruebas global e individual (excluyendo variables)

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
$H_1:$ Al menos un $\beta_i \neq 0$

Model Likelihood Ratio Test	
LR chi2	54.81
d.f.	4
Pr(> chi2)	<0.0001

AIC: 101.94

A un nivel de significancia del 0.05 se rechaza la hipótesis planteada por lo que concluimos que al menos una de las variables predictoras influye en el éxito o fracaso de adquirir el seguro.

Prueba de Wald

H_0= ß$_j$ = 0 (la información que se perdería al eliminar X_j en el siguiente caso no es significativa)

H1= ßj ≠ 0

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.880370	1.712946	-4.017	5.90e-05	***
Edad	0.125967	0.030386	4.146	3.39e-05	***
Hijos	0.420835	0.213740	1.969	0.04896	*
Empresa	-0.010351	0.004019	-2.575	0.01001	*
Automóvil	1.909898	0.634070	3.012	0.00259	**

Se aprecia que las variables sexo e ingresos, junto con el sexo para la categoría masculino no son significativos, por lo que no se perdería información si se eliminan del modelo

Validación del modelo

H0= El modelo ajustado es adecuado
H1= El modelo ajustado no es adecuado

```
        Hosmer and Lemeshow test (binary model)

data:  Seguros$Seguro, fitted(Seguros.m2)
X-squared = 7.7582, df = 8, p-value = 0.4574
```

No se rechaza la hipótesis planteada, por lo que se concluye que el modelo de regresión logística es adecuado o se ajusta a los datos (no hay diferencia significativa entre los valores observados y los que predice el modelo)

Bondad de ajuste y validez del modelo

```
$Pseudo.R.squared.for.model.vs.null
                                Pseudo.R.squared
McFadden                                0.373471
Cox and Snell (ML)                      0.400832
Nagelkerke (Cragg and Uhler)            0.537112
```

El pseudo-r cuadrado son respetables muestras de la variabilidad explicada por el modelo, y en ellas se observa que la Nagelkerke estima en un 53.71% tal variabilidad

- Cox y Snell: 40.08 % de la variabilidad total de la variable dependiente (respuesta) es explicado por las variables predictoras y el ruido o error (59.02%) no es explicado.

Formulación del modelo

$$Prob(Seguro = 1) = \frac{e^{-6.8804+0.1259*Edad+0.421*hijos-\ 0.0104*Empresa+1.9098*automóvil}}{1+e^{-6.8804+0.1259*Edad+0.421*\ hijos-\ 0.0104*Empresa+1.9098*automóvil}}$$

$\hat{\beta}_0 = -6.8804$, es el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ cuando las otras variables toman el valor de cero.

$\hat{\beta}_1 = 0.1259$, el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ se incrementa en 0.1259 cuando la edad de la persona aumenta en un año y las demás variables permanecen constantes.

$\hat{\beta}_2 = 0.421$, el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ se incrementa en 0.0421 cuando el número de hijo aumenta en 1 y las demás variables permanecen constantes.

$\hat{\beta}_3 = -0.0104$, el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ disminuye en 0.0104 cuando los años de la persona en la empresa aumenta en un año y las demás variables permanecen constantes.

$\hat{\beta}_4 = 1.9098$, el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ se incrementa en 1.9098 cuando el la persona cuenta con automóvil, y las demás variables permanecen constantes.

Interpretando los OR

	OR	2.5 %	97.5 %
(Intercept)	0.001027764	0.0000357957	0.02950908
Edad	1.134245216	1.0686662831	1.20384841
Hijos	1.523233451	1.0019148097	2.31580582
Empresa	0.989702729	0.9819374823	0.99752938
AutomóvilSí	6.752402432	1.9486635142	23.39805629

$e^{\widehat{\beta}_2} = 1.135$, el Resultado de adquirir un seguro como éxito, es 1.135 más ventajoso (probable) cuando la edad de la persona aumenta en un año.

$e^{\widehat{\beta}_4} = 6.511$, el Resultado de adquirir un seguro como éxito, es 6.7524 más ventajoso (probable) cuando la persona si cuenta con automóvil.

Gráfica de densidad

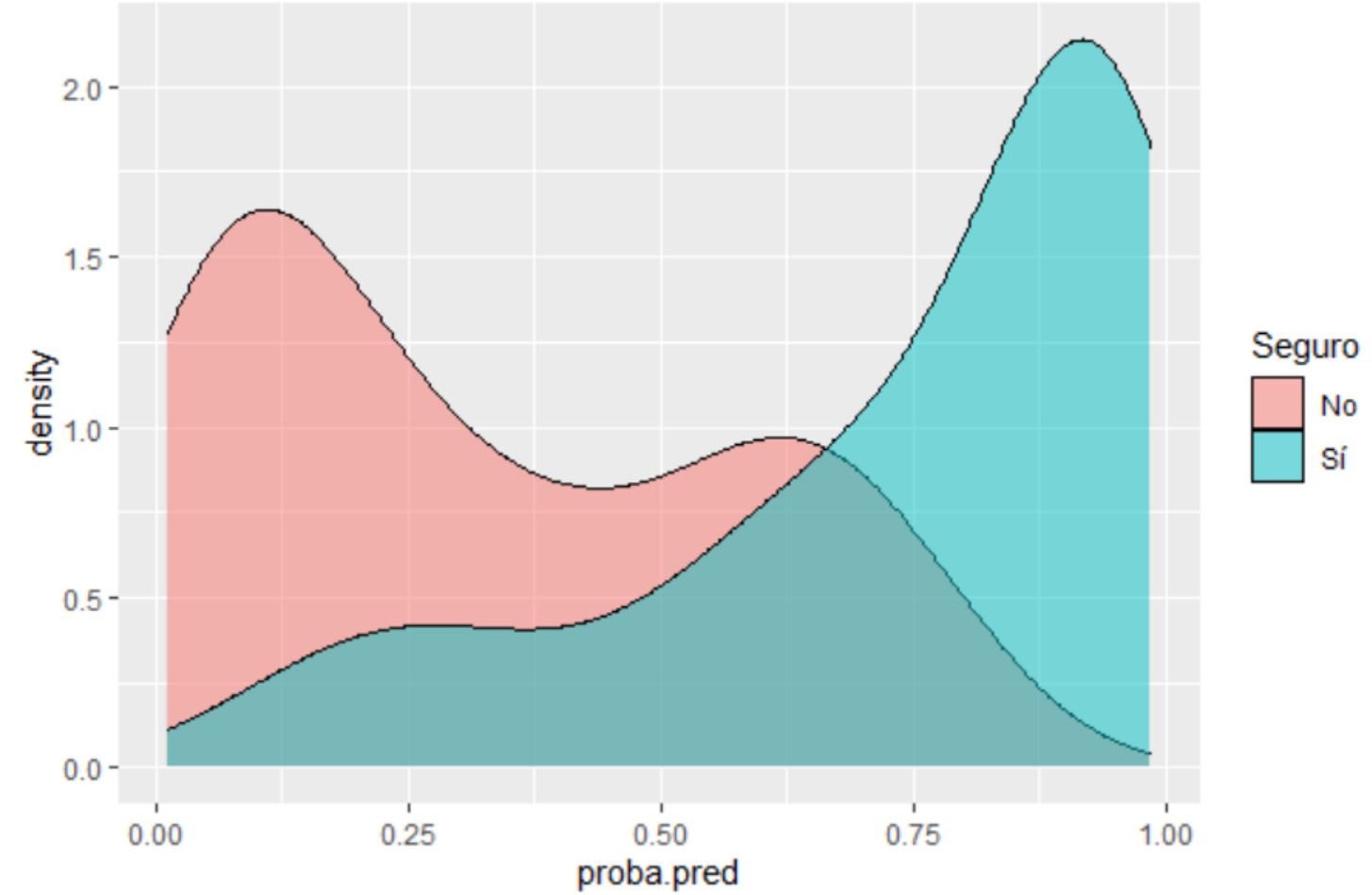


Tabla de clasificación correcta

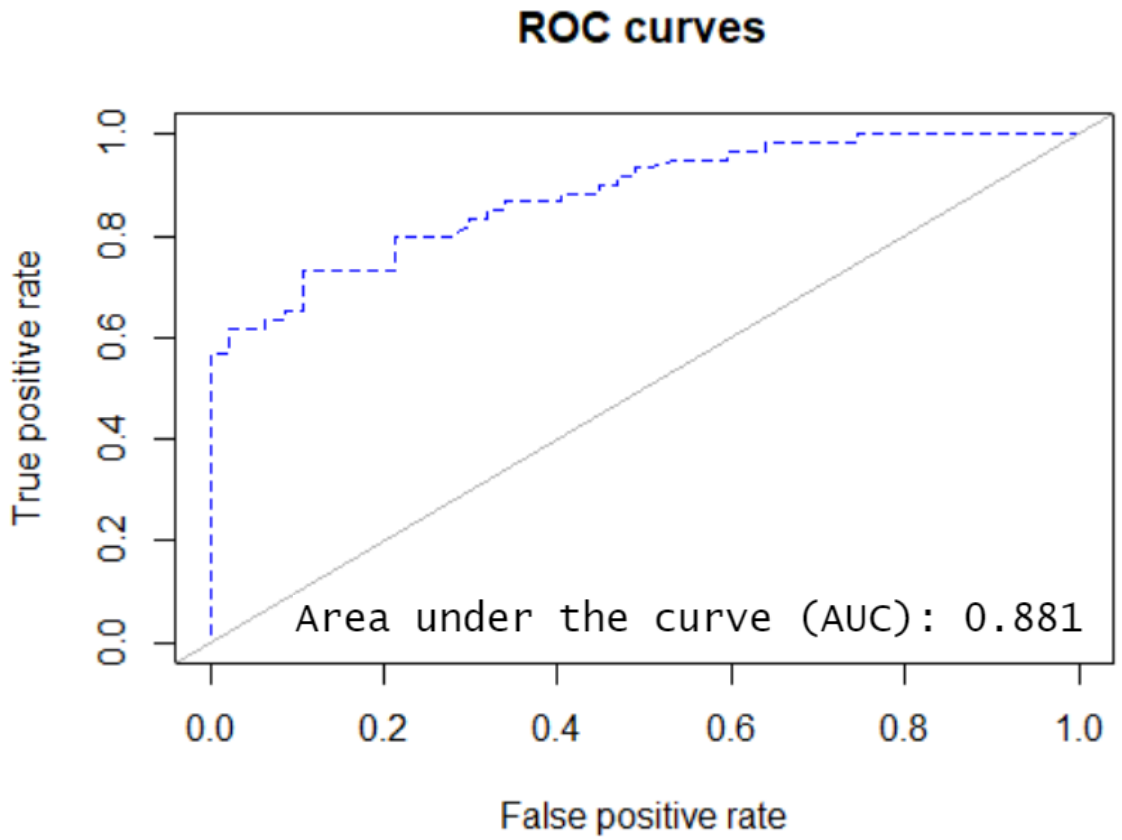
Tabla de clasificación

	Reference	
Prediction	No	Sí
No	32	10
Sí	15	50

Accuracy	0.7664
Error	0.2336
Sensibilidad	0.8333
Especificidad	0.6809

- El modelo clasifica correctamente a los clientes en un 76.635%
- La tasa de error del modelo es del 23.36%
- La probabilidad de clasificar correctamente a los que adquieren el seguro es de 0.8666.
- La probabilidad de clasificar correctamente a los que no adquieren el seguro es de 0.6809

Tabla de clasificación correcta



AUC ROC	Conclusión
0.5	Inaceptable
0.6 - 0.7	Pobre
0.71 - 0.8	Aceptable
0.81 - 0.9	Excelente (bueno)
> 0.9	Muy bueno

El AUC es 0.881, es decir el modelo es excelente, lo cual significa que existe un 88.1% de probabilidad de que el diagnóstico realizado a un cliente que adquiere el seguro sea más correcto que el de un cliente que no adquiere el seguro.

Prediciendo

Predecir si la persona adquirirá el seguro cuando es de sexo masculino, tiene 44 años, 3 hijos, tiene ingresos anuales de 1821 dólares, tiene 90 meses en la empresa, y tiene automóvil

$$ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 Edad + \beta_2 hijos + \beta_3 Empr + \beta_4 Auto(si)$$

$$ln\left[\frac{p}{1-p}\right] = -6.88 + 0.126 * 44 + 0.420 * 3 - 0.01 * 90 + 1.909 * 1 = 0.9030419$$

$$P(Seguro = Si) = \frac{e^{\alpha+\beta_k X_{ki}}}{1+e^{\alpha+\beta_k X_{ki}}} = \frac{e^{0.90300419}}{1+e^{0.90300419}} = 0.7116$$

$$odds = \frac{p}{1-p} = \frac{0.7116}{1-0.7116} = 2.467$$

- Como la probabilidad 0.71116 >0.5, entonces es más probable que el cliente con las características dadas si adquiera el seguro.
- Para persona con las siguientes características (tiene 44 años, 3 hijos, tiene 90 meses en la empresa, y tiene automóvil), hay 2.467 veces más probabilidad de adquirir el seguro a no adquirirlo.