Análisis Predictivo de Riesgo de Crédito: Estudio de la Base de Datos de Crédito Alemán (*German Credit*)

Ignacio Fernández Sánchez-Pascuala

Universidad Complutense Madrid & Universidad Politécnica Madrid

29 de abril de 2024

Contenido

Introducción

- Para minimizar las pérdidas, el banco necesita una regla de decisión con respecto a a quién otorgar la aprobación de un préstamo.
- Se consideran los perfiles demográficos y socioeconómicos de los solicitantes antes de tomar una decisión.
- *German Credit*: Información sobre 20 variables y la clasificación de si un solicitante es considerado un riesgo de crédito Bueno o Malo para 1000 solicitantes.

Contenido

Base de Datos Inicial

	Creditability	Account.Balance	Duration.of.Credit..month.	Payment.Status.of.Previous.Credit	Purpose	Credit.Amount	Value.Savings.Stocks
	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	1	1	18	4	2	1049	1
2	1	1	9	4	0	2799	1
3	1	2	12	2	9	841	2
4	1	1	12	4	0	2122	1
5	1	1	12	4	0	2171	1
6	1	1	10	4	0	2241	1

Limpieza/Transformaciones de Datos Realizadas:

- Valores numéricos a la categoría que representan.
- Combinación de categorías.
- Cambio nombre variables.

Base de Datos Limpia

Creditability	Account_Balance	Duration	Previous_Credit	Purpose	Amount	Savings	Employment_Length	Instalment	Sex/Martial	Guarantors
<fct>	<fct>	<int>	<fct>	<fct>	<int>	<fct>	<fct>	<fct>	<fct>	<fct>
Good	No account	18	No Problems	Used Car	1049	None	< 1 Year/Unemployed	< 20%	Male Divorced/Single	None
Good	No account	9	No Problems	Other	2799	None	[1,4)	(25%,35%)	Male Married/Widowed	None
Good	None	12	Paid Up	Other	841	< 100 DM	[4,7)	(25%,35%)	Male Divorced/Single	None
Good	No account	12	No Problems	Other	2122	None	[1,4)	[20%,25%)	Male Married/Widowed	None

Duration_Adress	Valuable_Asset	Age	Concurrent_Credits	Housing	Num_Credits	Occupation	Num_Dependents	Telephone
<fct>	<fct>	<int>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>
> 7	Car	21	None	Free	1	Skilled	3 or More	No
[1,4)	None	36	None	Free	> 1	Skilled	Less than 3	No
> 7	None	23	None	Free	1	Unskilled Permanent Resident	3 or More	No
[1,4)	None	39	None	Free	> 1	Unskilled Permanent Resident	Less than 3	No

Descripción de variables

- **Creditability**- Binaria. Confianza en la devolución del crédito: Malo y Bueno.
- **Duration**- Continua. Duración del préstamo en meses.
- **Amount**- Continua. Cantidad de crédito solicitada.
- **Age**- Continua. Edad del cliente.
- **Purpose**- Categórica. Propósito del crédito: Coche Nuevo, Coche Usado, Relacionado con el Hogar y Otro.
- **Housing**- Categórica. Vivienda del solicitante: Gratis, Alquilada y Propia.
- **Telephone**- Binaria. Disponibilidad de teléfono del cliente: Sí y No.
- **Occupation**- Categórica. Ocupación del solicitante: Desempleado, No Cualificado, Cualificado y Ejecutivo.
- **Num_Credits**- Binaria. Número de créditos en este banco: 1 y Más de 1.
- **Employment_Length**- Categórica. Duración del empleo actual: Menos de 1 año/Desempleado, [1,4), [4,7) y Más de 7.

- **Savings**- Categórica. Ahorros del cliente: Ninguno, Menos de 100 DM, [100, 1000] DM y Más de 1000 DM.
- **Previous_Credit**- Categórica. Estado de pago del crédito anterior: Algunos Problemas, Pagado y Sin Problemas.
- **Account_Balance**- Categórica. Saldo cuenta del solicitante: Sin cuenta, Ninguno y Algún saldo.
- **Instalment**- Categórica. Porcentaje de los ingresos disponibles destinados a pagos de cuotas/préstamos: Más del 35 %, (25 %, 35 %), [20 %, 25 %) y Menos del 20 %.
- **Sex/Martial**- Categórica. Género y estado civil del solicitante: Hombre Divorciado/Soltero, Hombre Casado/Viudo y Mujer.
- **Guarantors**- Binaria. Tipo de garante asociado con la solicitud de crédito: Ninguno y Sí.
- **Duration_Adress**- Categórica. Duración de residencia dirección actual: Menos de 1 año, [1, 4), [4, 7) y Más de 7.
- **Valuable_Asset**- Categórica. Activo más valioso: Ninguno, Coche, Seguro de Vida y Bienes Inmuebles.

- **Concurrent_Credits**- Binaria. Situación crediticia simultánea del solicitante en otras instituciones financieras: Otros Bancos o Tiendas y Ninguna.
- **Num_Dependents**- Binaria. Cantidad de personas económicamente dependientes del solicitante: 3 o Más y Menos de 3.

Variable Objetivo: Creditability
Distribución:

- Malo: 300
- Bueno: 700

Contenido

Cuadro: Tabla de contingencia entre Creditability y Account_Balance

Creditability	No account	None	Some Balance
Bad	135	105	60
Good	139	164	397

Cuadro: Tabla de contingencia entre Creditability y Previous_Credit

Creditability	Some Problems	Paid Up	No Problems
Bad	53	169	78
Good	36	361	303

Cuadro: Tabla de contingencia entre Creditability y Savings

Creditability	**None**	**$<$ 100 DM**	**[100,1000] DM**	**$>$ 1000 DM**
Bad	217	34	17	32
Good	386	69	94	151

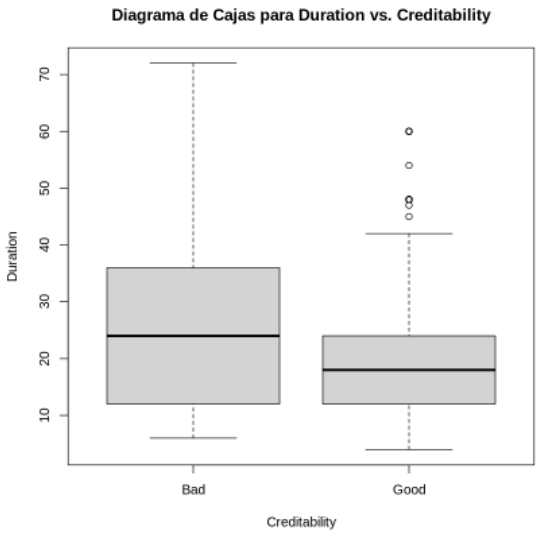
Cuadro: Tabla de contingencia entre Creditability y Employment_Length

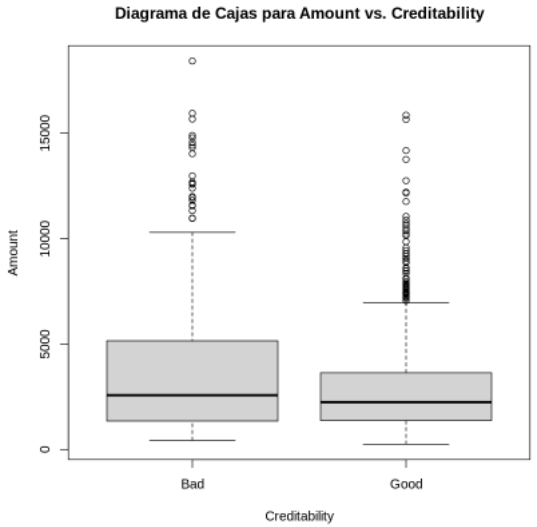
Creditability	**$<$ 1 Year/Unemployed**	**[1,4)**	**[4,7)**	**Above 7**
Bad	93	104	39	64
Good	141	235	135	189

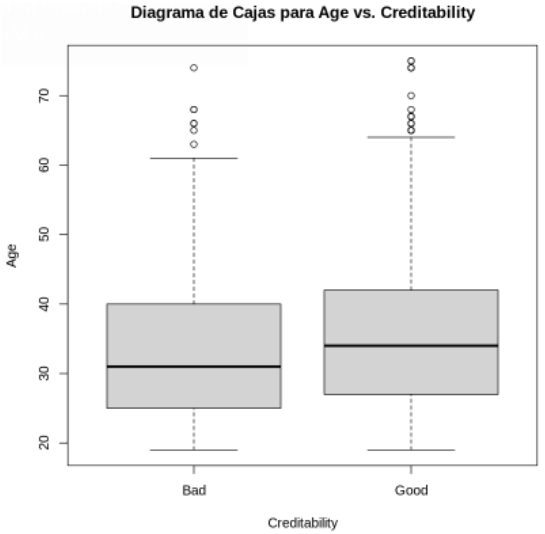
Cuadro: Tabla de contingencia entre Creditability y Sex/Martial

Creditability	**Male Divorced/Single**	**Male Married/Widowed**	**Female**
Bad	129	146	25
Good	231	402	67

Gráficos Cajas y Bigotes







Medidas de asociación:

- **Diferencia de proporciones (PD):** La diferencia entre la probabilidad de que ocurra un evento en un grupo y la probabilidad de que ocurra ese mismo evento en otro grupo.

$$DP = \pi_1 - \pi_2$$

  donde π_i es la probabilidad del suceso de referencia o de éxito para el grupo i.

- **Riesgo Relativo (RR):** La proporción de la probabilidad de que ocurra un evento en un grupo en comparación con la probabilidad de que ocurra ese mismo evento en otro grupo.

$$RR = \frac{\pi_1}{\pi_2}$$

- **Odds Ratio (OR):** La relación de que ocurra un evento en un grupo frente a la posibilidad de que ocurra ese mismo evento en otro grupo.

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Resultados de las Medidas de Asociación

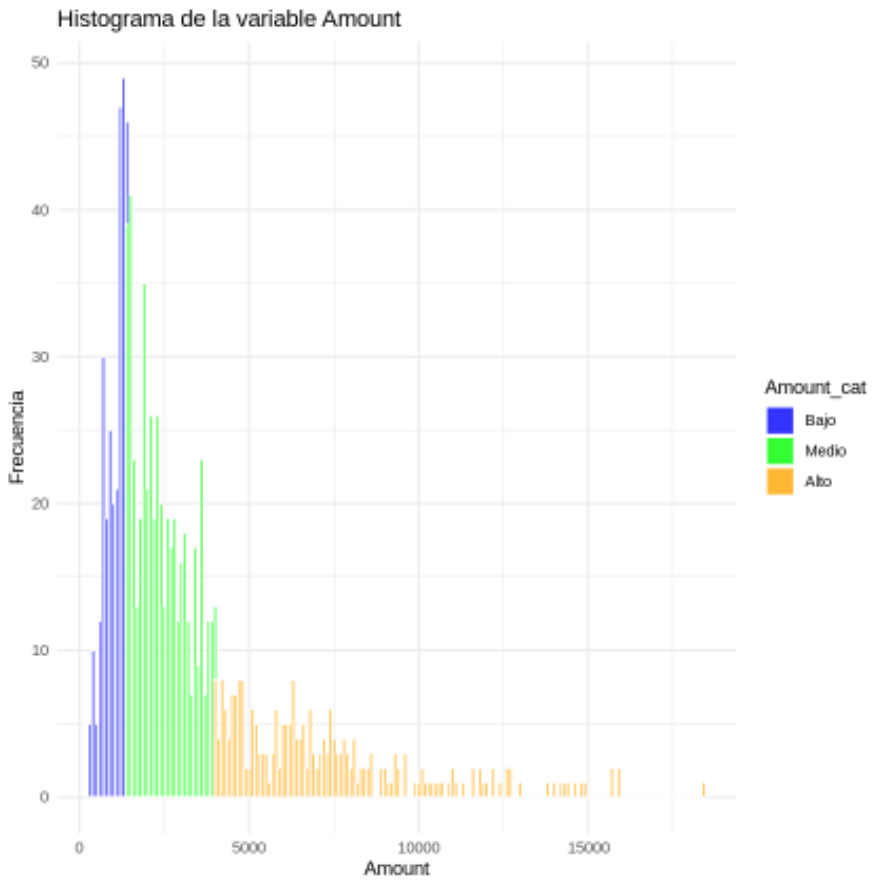
Variable	**PD**	**RR**	**OR**
Num_Dependents	0,0038 (−0,0745, 0,0821)	1,0129 (0,7784, 1,3179)	1,0184 (0,7002, 1,4813)
Telephone	0,0341 (−0,0234, 0,0915)	1,1218 (0,9217, 1,3652)	1,1774 (0,8919, 1,5543)
Num_Credits	0,0435 (−0,0147, 0,1017)	1,1596 (0,9468, 1,4201)	1,2333 (0,9280, 1,6389)
Concurrent_Credits	0,1334 (0,0564, 0,2104)	1,4848 (1,2087, 1,8240)	1,8198 (1,3078, 2,5322)
Guarantors	−0,0012 (−0,0991, 0,0967)	0,9961 (0,7195, 1,3789)	0,9944 (0,6244, 1,5835)

Razones de odds condicionales

Dado un factor condicional $Z = k$ sobre un evento, la razón de odds condicional (conditional odds ratios) está dada por la expresión:

$$\theta_{X,Y|Z=k} = \frac{n_{11}^k \cdot n_{22}^k}{n_{12}^k \cdot n_{21}^k}$$

- En nuestro caso, X es Creditability, Y es Concurrent_Credits y Z es Amount (categorizada).
- ¿Se acentúa el riesgo de incumplimiento de los clientes al tener créditos fuera de este banco conforme aumenta la cantidad del préstamo?
- Categorizamos la variable Amount en 3 categorías: Bajo, Medio y Alto.



Categoria_Z	Razones de odds condicionales
Bajo	2.844
Medio	1.431
Alto	1.671

Pruebas Estadísticas

- Test de Contingencia de Chi-cuadrado (χ^2):
  - Utilizado para determinar asociación significativa entre variables categóricas.
  - H0: No hay asociación entre las variables.
  - H1: Hay asociación significativa entre las variables.
  - Fórmula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- Prueba T de Student:
  - Utilizada para comparar las medias de dos grupos.
  - En el contexto de variables continuas y una variable respuesta categórica.
  - H0: No hay diferencia significativa entre las medias de los dos grupos.
  - H1: Sí hay diferencia significativa entre las medias de los dos grupos.
  - Fórmula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Resultados de las Pruebas Estadísticas

Variable	P-Value
Account_Balance	0.00
Previous_Credit	0.00
Purpose	0.00
Savings	0.00
Employment_Length	0.00
Instalment	0.14
Sex/Martial	0.01
Guarantors	1.00
Duration_Address	0.86
Valuable_Asset	0.00
Concurrent_Credits	0.00
Housing	0.00
Num_Credits	0.17
Occupation	0.60
Num_Dependents	1.00
Telephone	0.28

Variable	Media_Good	Media_Bad	P-Value
Duration	19.20714	24.860	2.404081e-10
Amount	2985.44286	3938.127	2.477103e-05
Age	36.22000	33.960	3.778175e-03

Cuadro: Pruebas T-test Variables continuas

Contenido

Objetivo

- Crear un modelo predictivo para determinar la probabilidad de devolución de créditos.
- Utilizar datos históricos de clientes para entrenar el modelo.
- Aplicar el modelo a nuevos clientes para prever su riesgo crediticio.
- La variable respuesta es Creditability.
- Evaluar el rendimiento del modelo mediante datos de entrenamiento y prueba.

La regresión logística es un modelo estadístico que modela la probabilidad de una variable binaria como una función lineal de una o más variables independientes:

$$P(Y = 1|X_1, X_2, \ldots, X_p) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}$$

Donde:

- $\beta_0, \beta_1, \ldots, \beta_p$ son los parámetros del modelo que se estiman durante el proceso de ajuste.

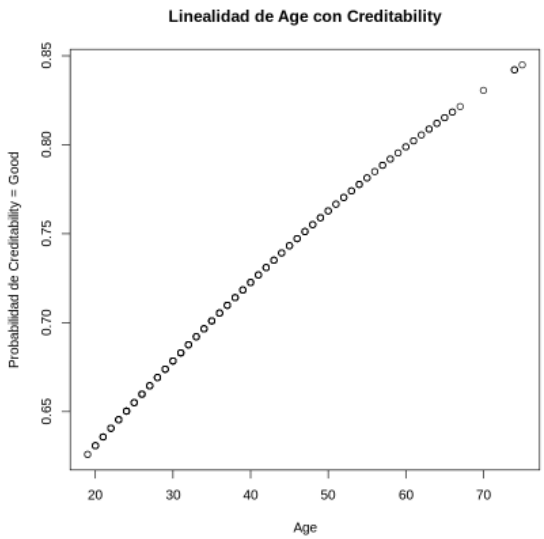
Separación de Datos y Selección de Variables

- Separación estratificada del conjunto de datos en entrenamiento (70 %) y prueba (30 %).
- Se consideran las variables con asociación significativa a la variable respuesta en pruebas anteriores.
- **Variables categóricas:** Account_Balance, Previous_Credit, Purpose, Savings, Employment_Length, Sex/Martial, Housing, Concurrent_Credits, Instalment y Valuable_Asset.
- **Variables continuas:** Age, Duration y Amount.

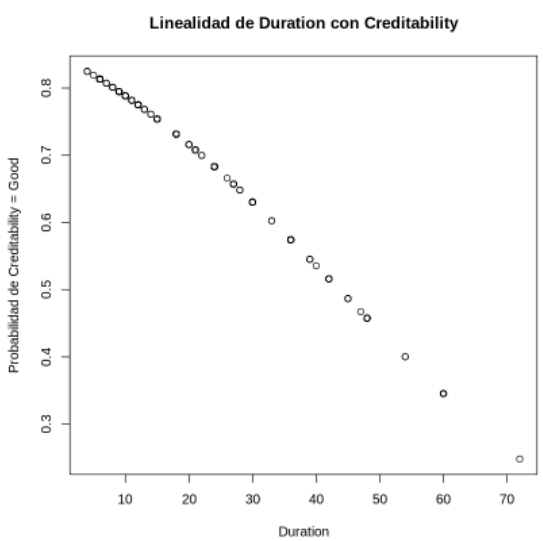
Variables continuas

- Se ajustaron modelos separados para las variables continuas respecto a la variable objetivo.
- Se evaluó la significancia de cada modelo mediante el p-valor del test de Wald.

P-valor para Age : 0.007166731



P-valor para Duration : 6.386738e-09



P-valor para Amount : 7.672581e-06

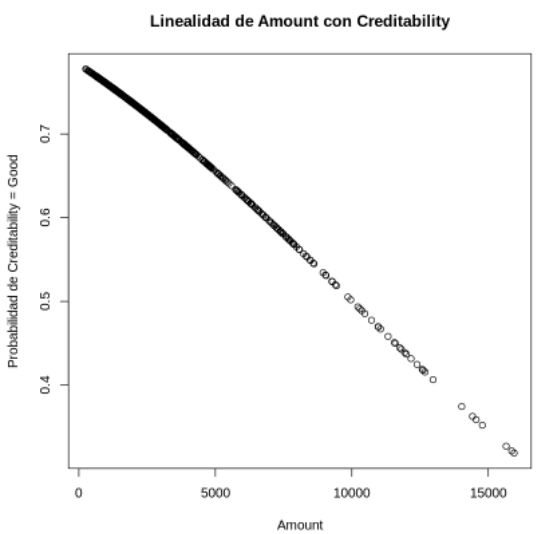
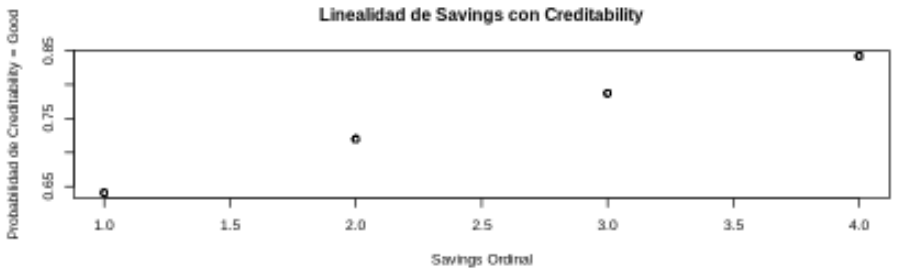


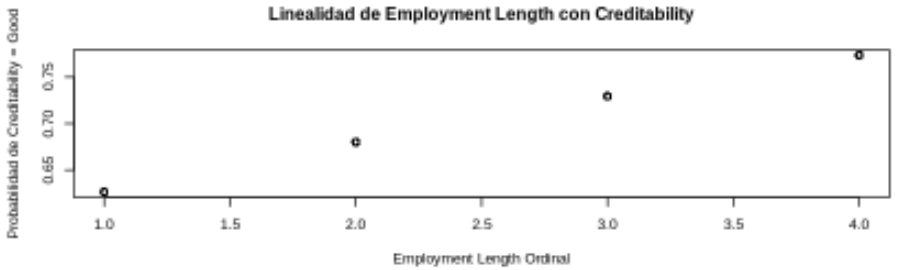
Figura: Modelos de regresión logística para las variables **Age**, **Duration** y **Amount**

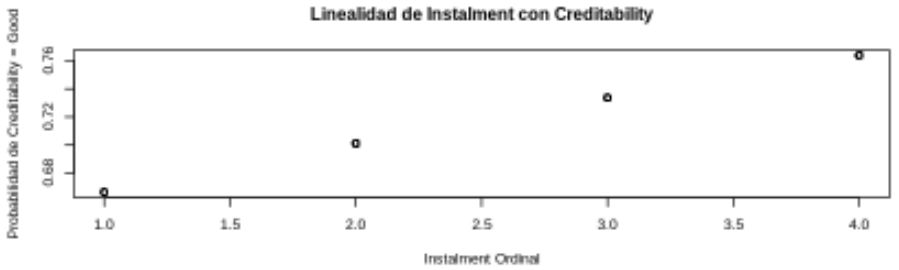
Variables Ordinales

- Variables Savings, Employment_Length e Instalment podrían ser ordinales.
- Se evalúa si existe una relación lineal entre sus categorías codificadas por su orden y la variable respuesta.
- Se entrena un modelo de regresión logística para evaluar la significancia de esta relación.

```
P-valor Savings: 5.413247e-06
P-valor Employment Length: 0.002182612
P-valor Instalment: 0.0361554
```







Modelo Inicial

- Método stepwise regression para entrenar el modelo.
- Se ajustan modelos con todas las combinaciones de variables y se comparan utilizando el Criterio de Akaike (AIC).
- Uso función glm de R para ajustar los modelos.

1. Entrenamos el modelo inicial con todas las variables seleccionadas ($AIC = 667{,}82$):

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.188e+00  6.941e-01  -4.593 4.36e-06 ***
Account_BalanceNone            5.490e-01  2.496e-01   2.199 0.027849 *
Account_BalanceSome Balance    1.832e+00  2.567e-01   7.135 9.65e-13 ***
Previous_CreditPaid Up         1.152e+00  3.786e-01   3.042 0.002352 **
Previous_CreditNo Problems     1.435e+00  3.862e-01   3.715 0.000203 ***
PurposeNew Car                 1.684e+00  4.159e-01   4.050 5.13e-05 ***
PurposeUsed Car                5.697e-01  2.851e-01   1.998 0.045719 *
PurposeHome Related            5.409e-01  2.468e-01   2.192 0.028411 *
`Sex/Martial`Male Married/Widowed 5.139e-01  2.239e-01   2.295 0.021739 *
`Sex/Martial`Female            2.124e-01  3.784e-01   0.561 0.574539
HousingRented                  5.494e-01  2.637e-01   2.083 0.037223 *
HousingOwned                   3.316e-01  5.255e-01   0.631 0.527949
Concurrent_CreditsNone         4.811e-01  2.655e-01   1.812 0.069999 .
Valuable_AssetCar             -5.536e-01  2.955e-01  -1.873 0.061017 .
Valuable_AssetLife Insurance  -6.295e-01  2.757e-01  -2.283 0.022410 *
Valuable_AssetReal Estate     -1.120e+00  4.574e-01  -2.449 0.014306 *
Age                            1.930e-02  1.013e-02   1.904 0.056897 .
Duration                      -2.613e-02  1.064e-02  -2.457 0.014020 *
Amount                        -1.069e-04  4.987e-05  -2.143 0.032120 *
Savings_ord                    2.967e-01  9.453e-02   3.139 0.001696 **
Employment_Length_ord          1.335e-01  1.009e-01   1.323 0.185797
Instalment_ord                 2.395e-01  1.013e-01   2.364 0.018058 *
```

Modelo final

2. Aplicamos el método Stepwise:

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.039e+00  6.858e-01  -4.431 9.37e-06 ***
Account_BalanceNone            5.324e-01  2.487e-01   2.141 0.032269 *
Account_BalanceSome Balance    1.841e+00  2.566e-01   7.174 7.29e-13 ***
Previous_CreditPaid Up         1.157e+00  3.767e-01   3.073 0.002122 **
Previous_CreditNo Problems     1.465e+00  3.840e-01   3.816 0.000136 ***
PurposeNew Car                 1.706e+00  4.179e-01   4.083 4.44e-05 ***
PurposeUsed Car                5.439e-01  2.838e-01   1.917 0.055253 .
PurposeHome Related            5.313e-01  2.458e-01   2.162 0.030639 *
`Sex/Martial`Male Married/Widowed 5.561e-01  2.212e-01   2.513 0.011954 *
`Sex/Martial`Female            2.173e-01  3.787e-01   0.574 0.566185
HousingRented                  5.408e-01  2.633e-01   2.054 0.040009 *
HousingOwned                   3.223e-01  5.253e-01   0.613 0.539567
Concurrent_CreditsNone         5.034e-01  2.644e-01   1.904 0.056865 .
Valuable_AssetCar             -5.802e-01  2.951e-01  -1.966 0.049245 *
Valuable_AssetLife Insurance  -6.269e-01  2.757e-01  -2.274 0.022994 *
Valuable_AssetReal Estate     -1.119e+00  4.587e-01  -2.439 0.014735 *
Age                            2.267e-02  9.891e-03   2.292 0.021928 *
Duration                      -2.563e-02  1.063e-02  -2.412 0.015860 *
Amount                        -1.066e-04  4.992e-05  -2.135 0.032775 *
Savings_ord                    3.078e-01  9.406e-02   3.273 0.001064 **
Instalment_ord                 2.316e-01  1.011e-01   2.290 0.022008 *
```

Se obtiene un AIC de 667.58.

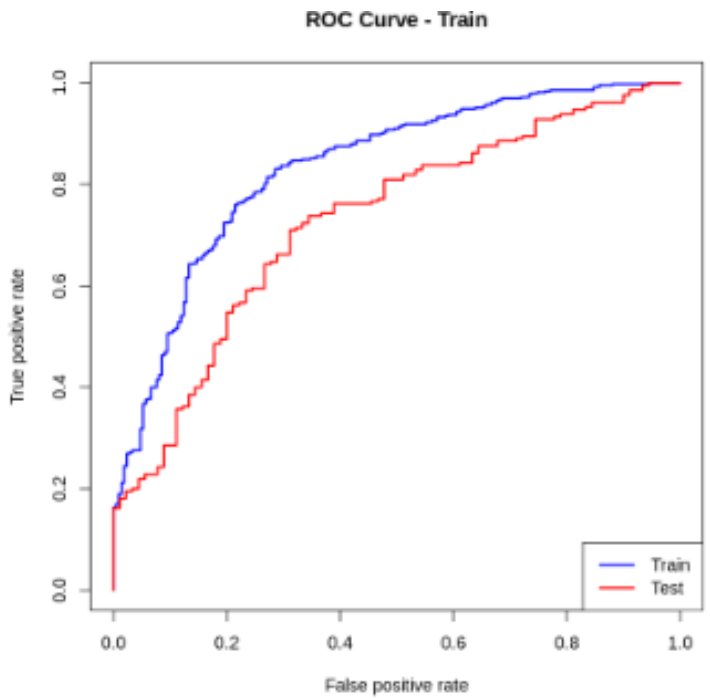
Modelo final

Odds Ratio modelo final:

(Intercept)	Account_BalanceNone
0.04787629	1.70301763
Account_BalanceSome Balance	Previous_CreditPaid Up
6.30333247	3.18159353
Previous_CreditNo Problems	PurposeNew Car
4.32842281	5.50919704
PurposeUsed Car	PurposeHome Related
1.72279260	1.70114924
`Sex/Martial`Male Married/Widowed	`Sex/Martial`Female
1.74381260	1.24268591
HousingRented	HousingOwned
1.71729542	1.38025102
Concurrent_CreditsNone	Valuable_AssetCar
1.65436827	0.55977816
Valuable_AssetLife Insurance	Valuable_AssetReal Estate
0.53423346	0.32667069
Age	Duration
1.02292613	0.97469378
Amount	Savings_ord
0.99989344	1.36049296
Instalment_ord	
1.26055817	

Evaluación del modelo

- **Métricas usadas:** AUC, precisión, exactitud y recall (Train y Test).



Métrica	Train	Test
AUC	0.833	0.727
Precision	0.897	0.828
Accuracy	0.787	0.72
Recall	0.816	0.783

Figura: Curvas ROC Train y Test

¡Gracias por su atención!

¿Alguna pregunta?