Construcción de un modelo de Credit Scoring

Instructora: Sherly Tarazona

Trayectoria profesional

Liderazgo

caja arequipa

Rol más técnico

YANBAL

BANCO PICHINCHA

SUNAT

Prácticas pre profesionales

Innova Schools

BCP

experian

Universidad-Pregrado

Universidad-Posgrado

unir UNIVERSIDAD INTERNACIONAL DE LA RIOJA

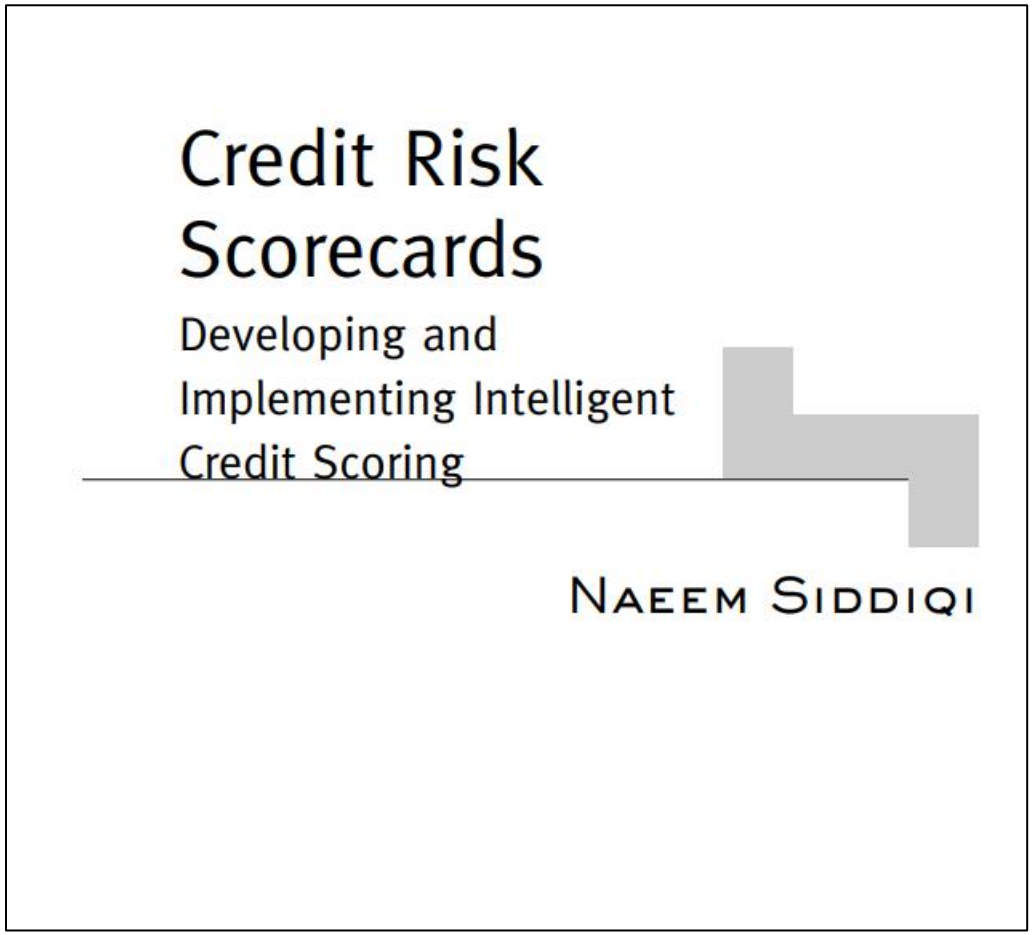
Voluntariado

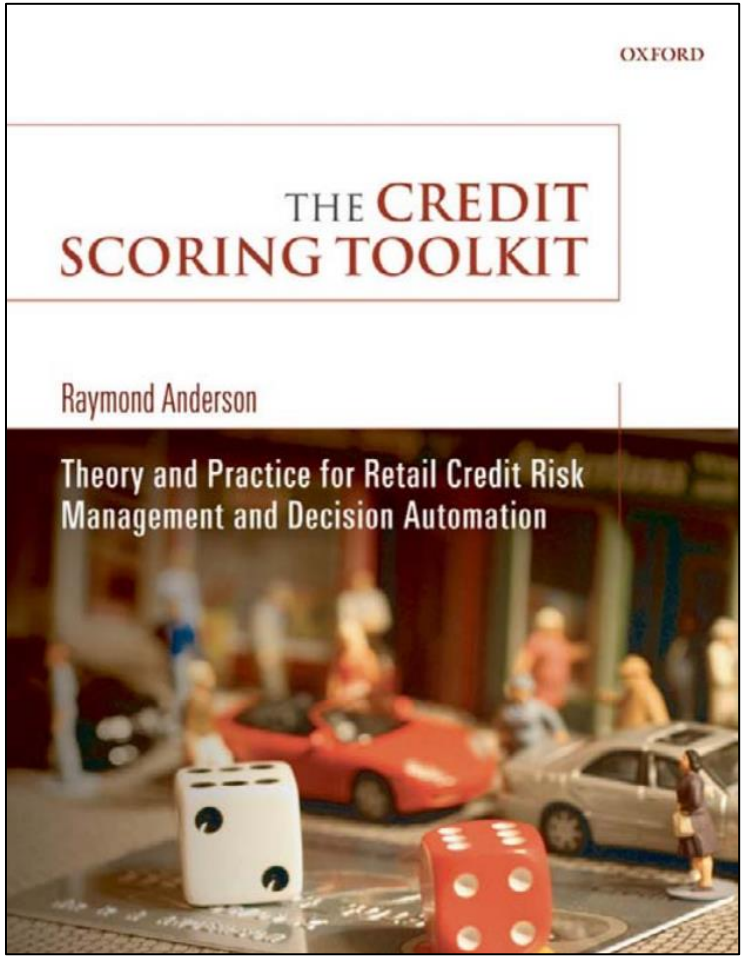
R-Ladies Lima

BILIOGRAFÍA BÁSICA

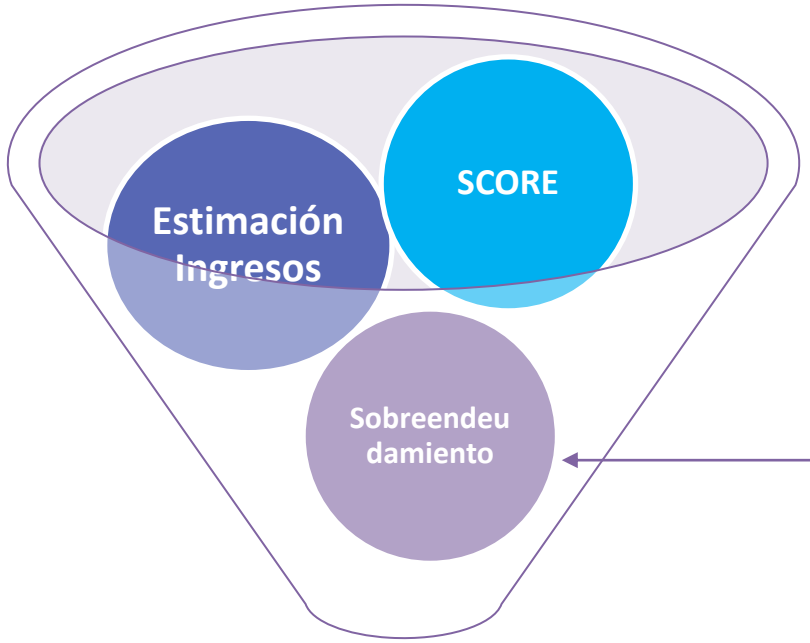
- Anderson, R(2007) *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation* . Oxford University Press
- Dalgaard, P (2008): Introductory Statistcs with R. Springer
- Mays,E and Niall Lynas (2011) Credit Scoring for Risk Managers: The Handbook for Lenders.Createspace (ISBN13: 9781450578967)
- Siddiqi, N. (2006): Credit Risk Scorecards. Devoloping and implementing Intelligent Credit Scoring. J Wiley & Sons
- Trueck, S, & Rachev, Svetlozar (2009): Rating Based Modeling of Credit Risk. Theory and Application of Migration Matrices. Elsevier
- William, G. (2011). Data Mining with Rattle and R, The art of Excavating Data for Knwledge Discovery. Springer.
- Wooldridge, J.M. (2016) . Introductory Econometrics. Amodern Approach. 6a Edicion Cengage Learning. (Cap 17)

Libros


Credit Risk
Scorecards
Developing and
Implementing Intelligent
Credit Scoring
Naeem Siddiqi


OXFORD
THE CREDIT
SCORING TOOLKIT
Raymond Anderson
Theory and Practice for Retail Credit Risk
Management and Decision Automation

APLICACIÓN DE MODELOS DE RIESGOS PARA NEGOCIOS CON EVALUACIÓN MASIVA



Endeudamiento

$$RCI = \frac{Cuota\ estimada}{Ingreso\ estimado}$$

Cuota estimada $=\sum$

- Cuota Préstamo Consumo
- Cuota Tarjeta de Crédito Consumo
- Cuota Deuda Hipotecaria
- Cuota Deuda Convenio
- Cuota Deuda MES
- Cuotas Otras y Contingentes

Cálculo de Oferta

Oferta Préstamo Personal	
Monto Préstamo	**4,000**
Plazo (meses)	24
TEA	40.00%
Cuota Mensual	**S/. 250.00**

TIPOS DE MODELOS CREDIT SCORING

Admisión

Score de originación

Score de Buró

Score de fraude

Seguimiento

- Score de comportamiento (Behavior)
- Modelo de Alertas Tempranas

Cobranzas

- Score de cobro temprano.
- Score de cobro intermedio
- Score de cobro tardío

Recuperación

Score de recupero

Credit Scoring: Valoración o Puntuación del riesgo de crédito

ESTIMACION DE LA PROBABILIDAD DE IMPAGO

Se le asigna a cada individuo una puntuación o una probabilidad de que sea impagado y en función de dicha probabilidad se le concede o no el crédito.

La puntuación suele ir de menor a mayor calidad crediticia

Scorecard o Tarjeta de Puntuación del Riesgo

Variable	Atributo	Puntuación
Edad	Menor < 23	63
Edad	23-28	76
Edad	28-34	79
Edad	34-46	85
Edad	46-51	94
Edad	51- Mayor	105
Tipo Tarjeta	AMEX, VISA, Sin TRJ	80
Tipo Tarjeta	MasterCard	99
Salario	Menor <600	85
Salario	600- 1200	81
Salario	1200- 2200	93
Salario	2200 > Mayor	99
Estado Civil	Casado	85
Estado Civil	Resto	78

Ejemplo aplicación

Campo	Valor	Puntos
ID	Manuel T.	
Núm Solicitud	12345678	
Edad	43	85
Tipo TRJ	AMEX	80
Salario	1350	93
Estado Civil	Casado	85
TOTAL		**343**

Se establece una puntuación mínima o ***Cutt-off***. Cuanto mayor puntuación mejor calidad crediticia. Hay que establecer un mínimo para aprobar

VENTAJAS DE LOS CREDIT SCORING

Los modelos de credit scoring se han implementado desde hace unos 60 años, siendo su objetivo el predecir los potenciales patrones medios de pago de los clientes (default), permitiendo de esta manera conseguir las siguientes ventajas:

- Resuelven problemas como la relación entre variables, el potencial discriminante de las variables y la relación entre niveles de riesgo
- Discriminación de clientes buenos y malos (reducción de subjetividad)
- Eficiencia de costos (reducción en los tiempos de respuesta)
- Pricing de las operaciones acorde al riesgo asumido

Pequeñas reducciones en el riesgo de la cartera (con importante suma de capital) significan enormes incrementos en la rentabilidad del negocio.

DESVENTAJAS DE LOS CREDIT SCORING

Los modelos de credit scoring también presentan ciertas limitaciones:

- Requiere datos sobre una gran cantidad de préstamos y de cada préstamo
- Supone que el futuro será como el pasado
- Son sistemas que se deterioran con el tiempo

TIPOS DE VARIABLES EN SCORING

- **Numéricas (continuas)**
  - Edad
  - N° Entidades Financieras
  - Monto de Deuda
  - Líneas de Crédito
  - Días de atraso

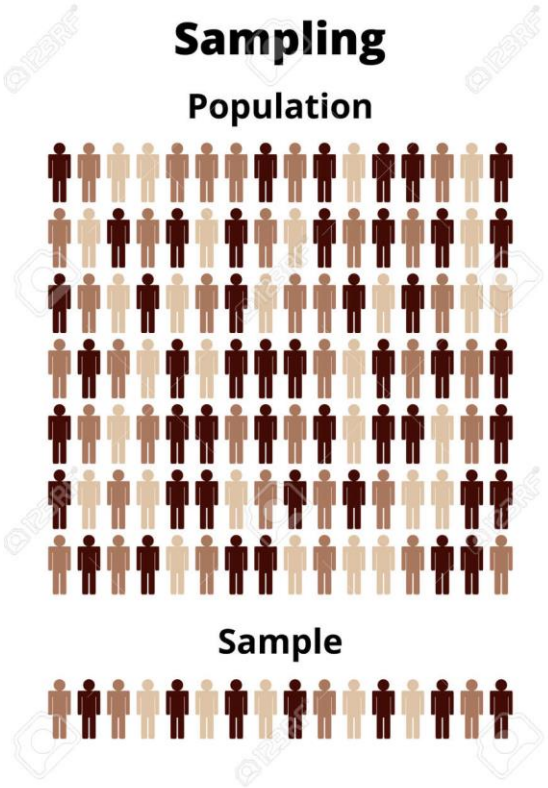
- **Categóricas (binaria o múltiple)**
  - Grado de instrucción
  - Ubicación Geográfica
  - Tipo de empleo
  - Estado civil
  - Default (0: Bueno, 1: Malo)

Nota: Se suelen crear diversos cálculos a partir de estas variables como máximos, mínimos, recencia, frecuencia, ratios, en el tiempo, es decir en los últimos 3m, 6m, 9m, 12m

SELECCIÓN DE LA MUESTRA

Muestreo Aleatorio Simple

- En esta técnica cada miembro de la población tiene la misma probabilidad de ser seleccionado como sujeto de análisis. Todo el proceso de selección de la muestra se realiza en un paso, en donde cada individuo es seleccionado independientemente de los otros miembros de la población.
- Por lo general de seleccionan muestras en tiempo (train y test) y fuera de tiempo (backtest).



Nota: Hoy en día debido al gran poder computacional de la nube, se suele trabajar con toda la población.

SELECCIÓN DE LA MUESTRA

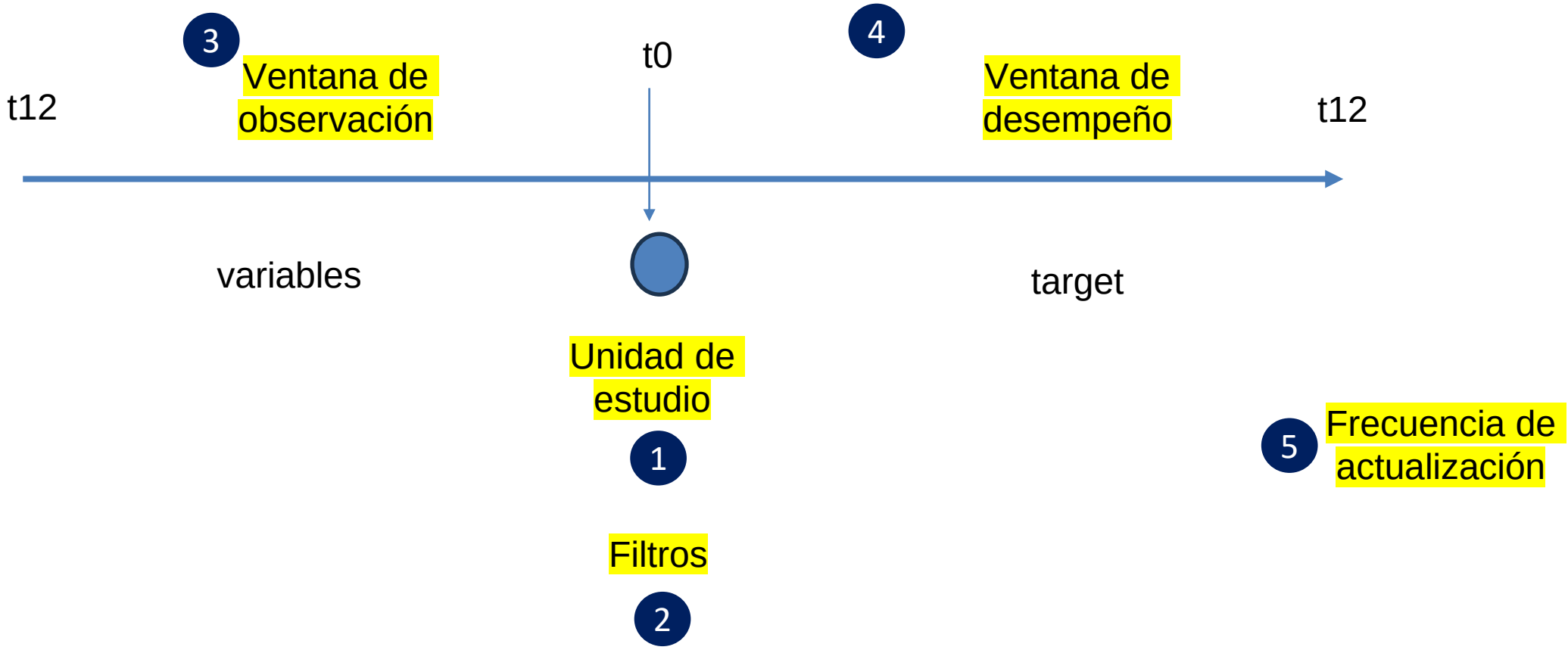
- Alternativamente se podría particionar la muestra de la siguiente forma:
  - Desarrollo: 70%
  - Validación: 30%

- Con la **muestra de desarrollo** se deberá realizar el análisis univariado antes de la estimación del modelo.

- Las **muestras de validación y de prueba** servirán para confirmar la capacidad predictiva del modelo, obtenida con la muestra de desarrollo.

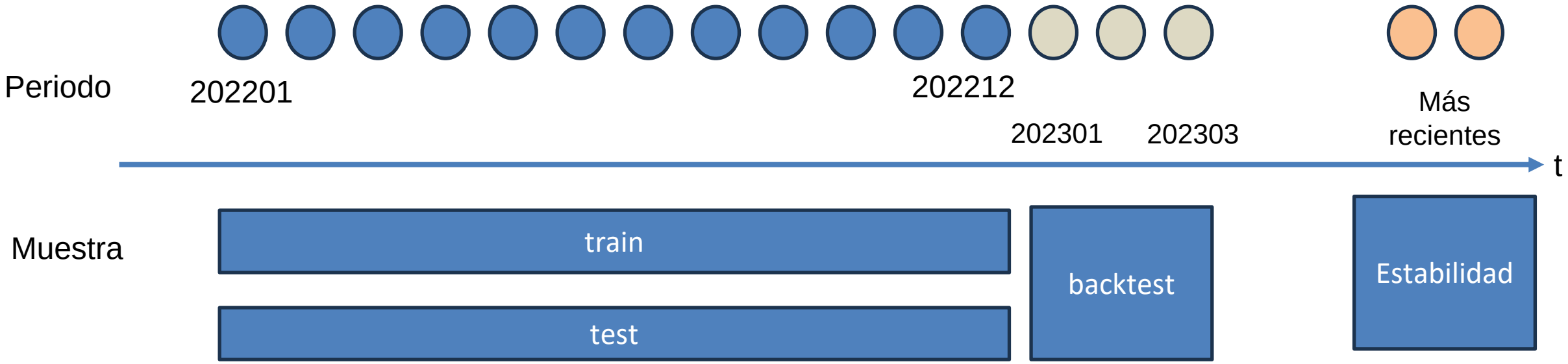
Diseño del modelo

Ejemplo:



Diseño del modelo

6 Elección de periodos de modelamiento



DEFINICIÓN DEL DEFAULT

Definiciones

Ejemplo: Variable objetivo de un Score de Buró.

- **Malo:** en alguno de los próximos 12 meses tuvo más de 30 días de atraso.
- **Bueno:** durante los 12 meses tuvo cero días de atraso.
- **Indeterminado:** en alguno de los próximos 12 meses tuvo entre 1 y 30 días de atraso.
- **Insuficiente:** no cuenta con información completa en los 12 meses.

ROLL RATES ANALYSIS

Días de atraso (t)	Días de atraso (t+12)					Total
	0 - 8	9 - 30	31 - 60	61 - 120	121 - 180	
0 - 8	52,187	1,350	1,260	1,575	3,929	60,301
9 - 30	701	73	40	87	815	1,716
31 - 60	320	7	25	39	1,193	1,584
61 - 120	66	1	2	7	731	807
121 - 180					1	1
Total	**53,274**	**1,431**	**1,327**	**1,708**	**6,669**	**64,409**

Deterioro
13%
55%
78%
91%

Pasado los 30 días de atraso, la mayoría de los créditos siguen avanzando a tramos más avanzados.

ESPECIFICACIÓN DE EXCLUSIONES

Las principales exclusiones realizadas a los modelos scoring son:

- Deudores que vienes con deterioro
- Deudores con poca información o insuficiente
- Deudores considerados outliers
- Entre otras que defina el negocio

VALORES PERDIDOS

- Los campos con valores perdidos *(missing values)* pueden ser o bien porque no fueron capturados, su recolección se descontinuó, no estaban disponibles, o no fueron completados por los solicitantes de crédito. También pueden existir valores con errores de tecleado *(mis-keyed values)* cuya inclusión en la base debe ser revisada.
- Mientras que algunas técnicas estadísticas tales como árboles de decisión son neutrales a los valores perdidos, la regresión logística requiere conjuntos de datos completos sin datos perdidos.

VALORES PERDIDOS

Hay cuatro formas principales para tratar los valores perdidos:

1. Excluir todos los datos con valores perdidos.
2. Excluir características o registros que tienen valores perdidos importantes del modelo (por ejemplo, más del 50%), especialmente si se espera que el nivel de falta continúe en el futuro.
3. Incluir características con valores perdidos en el scorecard. El "valor perdido" puede entonces ser tratado como un atributo independiente, agrupados, y se utiliza en la regresión como una entrada.
4. Imputar valores a valores perdidos utilizando técnicas estadísticas, según el tipo de variable: continua (media) o categórica (moda).

VALORES PERDIDOS

Ejemplos:

- **<u>Caso 1</u>: variable "Líneas de crédito"**

  Valores perdidos (nulos) corresponden a personas que no poseen tarjetas de crédito.

- **<u>Caso 2</u>: variable "Clasificación de riesgo en el sistema"**

  Un valor perdido (nulo) en esta variable correspondería a que en determinado mes la persona no registraba deudas y por lo tanto no se le asignó una clasificación de riesgo.

DETECCIÓN Y TRATAMIENTO DE OUTLIERS

- Los outliers (valores atípicos), se detectan principalmente en variables numéricas (no en categóricas).
- Una vez que haya identificado los casos atípicos, podrá examinarlos y determinar si debería incluirlos o no en los análisis.

PESO DE LA EVIDENCIA (WOE)

- Este indicador es una medida relativa que indica que tan discriminante es cada atributo de cada variable predictiva. Se mide como el LN del ratio de participación de buenos respecto de la participación de malos en el total de cada grupo.

$$WoE_i = Ln\left[\frac{\frac{Buenos_i}{Buenos\ Totales}}{\frac{Malos_i}{Malos\ Totales}}\right]$$

- Esta medida se utiliza para comparar cada atributo de cada variable predictiva, ya que no considera las unidades en las que fue medida cada variable.

VALOR DE LA INFORMACIÓN (IV)

- Este indicador es una medida relativa que indica que tan discriminante es cada atributo y cada variable predictiva. Su cálculo se basa en el indicador WoE.

$$IV_i = \left[\frac{Buenos_i}{Buenos\ Totales} - \frac{Malos_i}{Malos\ Totales}\right] * WoE_i$$

$$IV_j = \sum_{i=1}^{n} IV_i$$

Donde: j: variable, i: atributo

- Un valor más alto de IV, indica que la variable tiene más poder discriminante.

MEDICIÓN DEL PODER DISCRIMINANTE POR VARIABLE

Experiencia Laboral

N°	Atributo	Buenos	Malos	Total	PD	WoE	IV Atributo	IV
1	<= 11	3,452	860	4,312	20%	-0.36	0.01	0.01
2	12 - 15	3,373	963	4,336	22%	-0.50	0.02	0.02
3	16 - 20	3,078	793	3,871	20%	-0.39	0.01	0.03
4	21 - 25	3,049	725	3,774	19%	-0.31	0.01	0.04
5	26 - 32	3,846	877	4,723	19%	-0.27	0.00	0.04
6	33 - 37	3,233	670	3,903	17%	-0.18	0.00	0.04
7	38 - 44	3,487	658	4,145	16%	-0.08	0.00	0.04
8	45 - 50	3,166	616	3,782	16%	-0.11	0.00	0.04
9	51 - 59	3,498	643	4,141	16%	-0.06	0.00	0.04
10	60 - 68	3,348	551	3,899	14%	0.06	0.00	0.04
11	69 - 79	3,474	554	4,028	14%	0.09	0.00	0.04
12	80 - 93	3,387	564	3,951	14%	0.04	0.00	0.04
13	94 - 109	3,500	533	4,033	13%	0.13	0.00	0.05
14	110 - 130	3,758	558	4,316	13%	0.16	0.00	0.05
15	131 - 147	3,294	529	3,823	14%	0.08	0.00	0.05
16	148 - 169	3,649	498	4,147	12%	0.24	0.00	0.05
17	170 - 201	3,650	413	4,063	10%	0.43	0.01	0.06
18	202 - 251	3,646	391	4,037	10%	0.48	0.01	0.07
19	252 - 311	3,735	349	4,084	9%	0.62	0.02	0.08
20	312+	3,688	312	4,000	8%	0.72	0.02	0.10
	Total	69,311	12,057	81,368	15%			

IV de la variable

CATEGORIZACIÓN O TRAMEADO DE VARIABLES

- El proceso de optimización de variables categóricas parte de la agrupación de los atributos con PDs similares, otro criterio a utilizar podría ser WoE similares. Recién cuando el WoE de un atributo representa el doble (o un valor cercano) del WoE del atributo anterior se podría separar en otro grupo.
- Ubicar la categoría que agrupa los valores nulos según corresponda en base a la PD de este atributo.
- El proceso de optimización de variables continuas parte de la categorización de la variable en deciles y su posterior agrupación siguiendo los mismos criterios que en el caso de variables categóricas.

INTERACCIÓN DE VARIABLES

Las interacciones entre las variables se observarán en el árbol de clasificación.

También podría probarse utilizando tablas cruzadas entre 2 variables y luego añadiendo variables de manera que ayude a encontrar variables combinadas más predictivas.

En este tipo de modelos de scoring las variables combinadas que frecuentemente son seleccionadas son:

- Estado Civil vs. Sexo
- Rango de Edad vs. Sexo
- Lugar de Procedencia vs. Sector Económico
- Clasificación de Riesgo Sistema vs. N° Meses con Experiencia Crediticia

MEDICIÓN DEL PODER DISCRIMINANTE POR VARIABLE

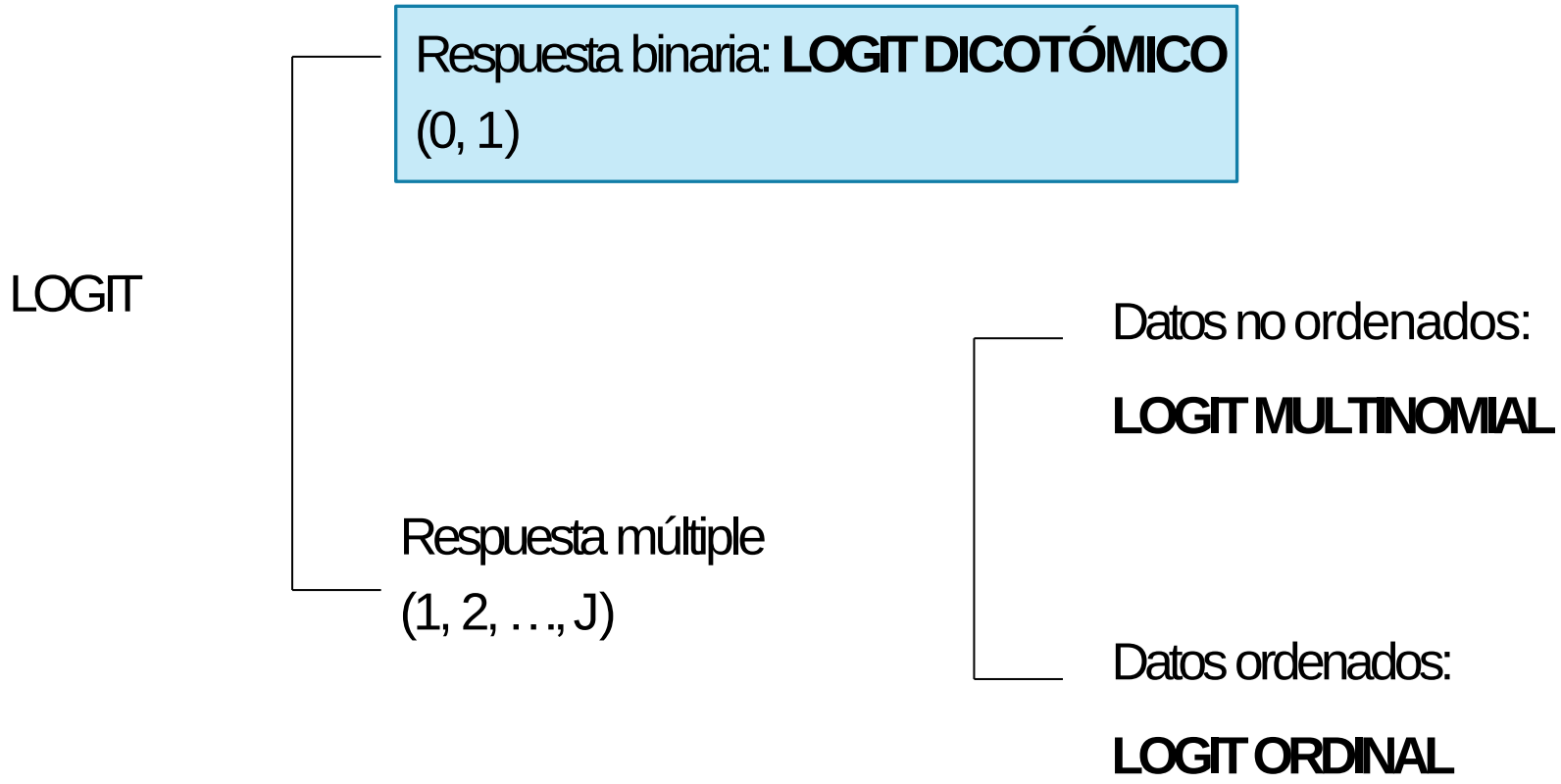
Experiencia Laboral

N°	Atributo	Buenos	Malos	Total	PD	WoE	IV Atributo	IV
1	<= 11	3,452	860	4,312	20%	-0.36	0.01	0.01
2	12 - 15	3,373	963	4,336	22%	-0.50	0.02	0.02
3	16 - 20	3,078	793	3,871	20%	-0.39	0.01	0.03
4	21 - 25	3,049	725	3,774	19%	-0.31	0.01	0.04
5	26 - 32	3,846	877	4,723	19%	-0.27	0.00	0.04
6	33 - 37	3,233	670	3,903	17%	-0.18	0.00	0.04
7	38 - 44	3,487	658	4,145	16%	-0.08	0.00	0.04
8	45 - 50	3,166	616	3,782	16%	-0.11	0.00	0.04
9	51 - 59	3,498	643	4,141	16%	-0.06	0.00	0.04
10	60 - 68	3,348	551	3,899	14%	0.06	0.00	0.04
11	69 - 79	3,474	554	4,028	14%	0.09	0.00	0.04
12	80 - 93	3,387	564	3,951	14%	0.04	0.00	0.04
13	94 - 109	3,500	533	4,033	13%	0.13	0.00	0.05
14	110 - 130	3,758	558	4,316	13%	0.16	0.00	0.05
15	131 - 147	3,294	529	3,823	14%	0.08	0.00	0.05
16	148 - 169	3,649	498	4,147	12%	0.24	0.00	0.05
17	170 - 201	3,650	413	4,063	10%	0.43	0.01	0.06
18	202 - 251	3,646	391	4,037	10%	0.48	0.01	0.07
19	252 - 311	3,735	349	4,084	9%	0.62	0.02	0.08
20	312+	3,688	312	4,000	8%	0.72	0.02	0.10
	Total	69,311	12,057	81,368	15%			

IV de la variable

REGRESIÓN LOGÍSTICA

- **Tipos de Modelos Logit**



ESTIMACIÓN DEL MODELO USANDO REGRESIÓN LOGÍSTICA

- Ecuación a estimar:

$$Prob(Y = Malo) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 Var1 + \beta_2 Var2 + \cdots + \beta_n VarN)}}$$

- Función de enlace: Logit, proviene de la familia de distribución binomial

ESTIMACIÓN PROBABILIDAD DE DEFAULT

- Se deben especificar las variables (dependiente e independientes).
- Identificar que variables con categóricas
- Luego de estimar la regresión, calcular la probabilidad de default.
- Finalmente evaluar la capacidad predictiva usamos, el KS, Área Curva ROC, Coeficiente Gini, Matriz de error o matriz de confusión.

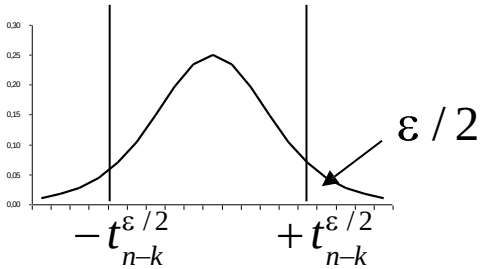
MEDIDAS DE AJUSTE INDIVIDUAL

INDIVIDUAL: Contraste de hipótesis

1. $H_0 : \beta = 0$
2. Estadístico de contraste Wald

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

3. Regla de decisión



Acepto H_0 si:

Valor de estadístico Wald < $\left|t_{n-k}^{\varepsilon/2}\right|$

Niv. sig. > ε

MEDIDAS DE BONDAD DE AJUSTE

CONJUNTA

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

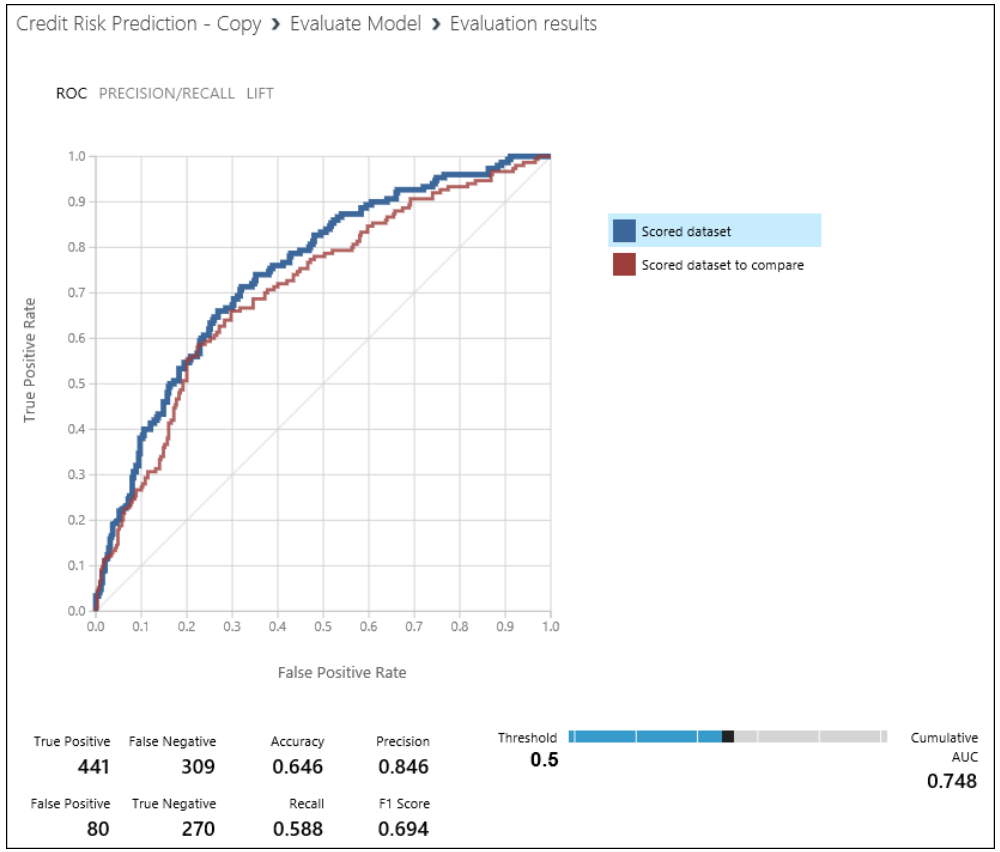
2. Razón de Verosimilitud = $X^2 = -2\log L(reducido) - \left[-2\log L(completo)\right]$

3. Porcentaje de aciertos: a través de un punto de corte

EVALUACIÓN DE MODELOS CREDIT SCORING

Existen diferentes indicadores para evaluar y comparar la capacidad predictiva de los modelos scorecard.

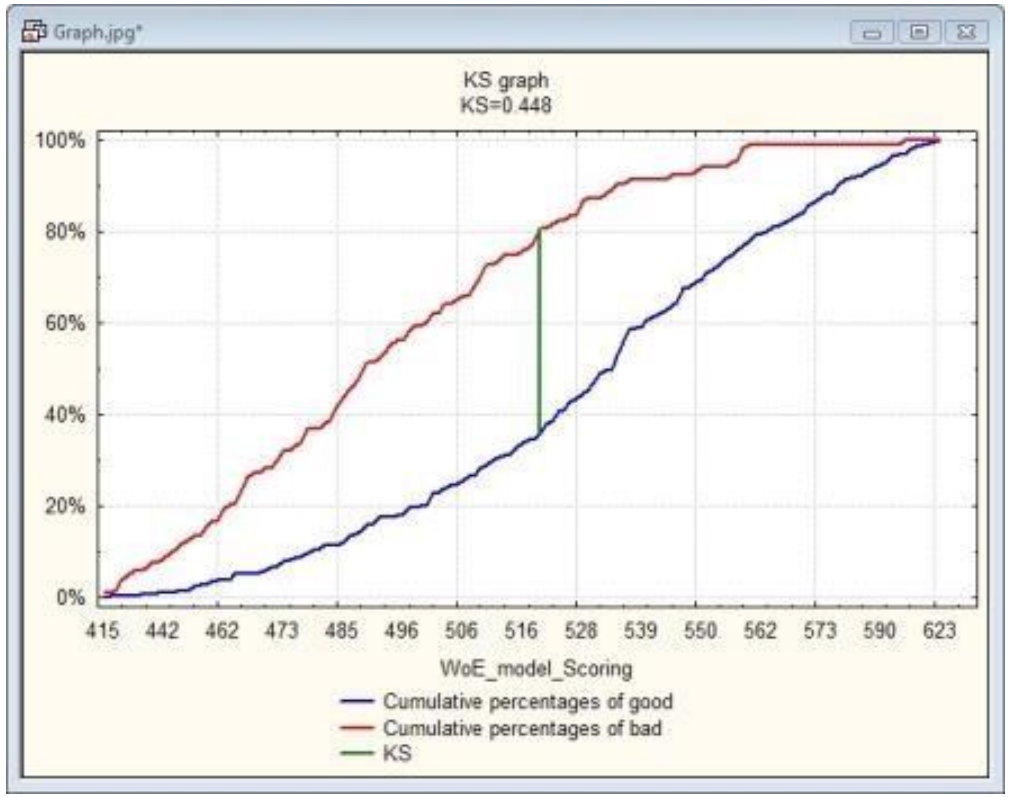
- **Curva ROC**



EVALUACIÓN DE MODELOS CREDIT SCORING

Existen diferentes indicadores para evaluar y comparar la capacidad predictiva de los modelos scorecard.

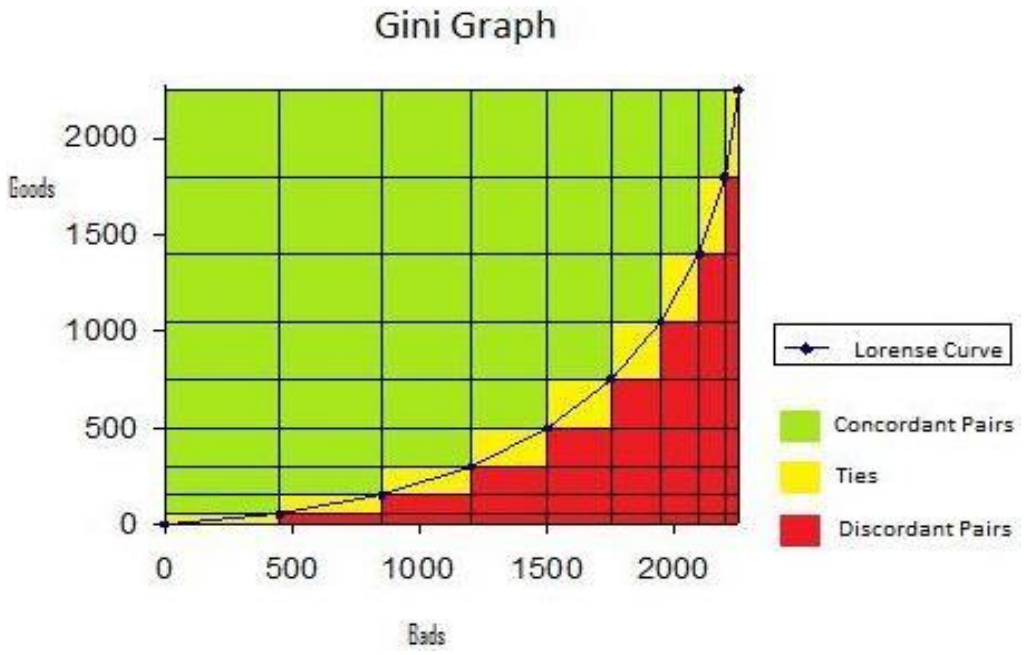
- **Curva ROC**
- **KS**



EVALUACIÓN DE MODELOS CREDIT SCORING

Existen diferentes indicadores para evaluar y comparar la capacidad predictiva de los modelos scorecard.

- **Curva ROC**
- **KS**
- **Gini**



Equivalencia Gini y Área Curva ROC

Gini = 2(AUC) - 1

EVALUACIÓN DE MODELOS CREDIT SCORING

Existen diferentes indicadores para evaluar y comparar la capacidad predictiva de los modelos scorecard.

- **Curva ROC**
- **KS**
- **Gini**
- **Matriz de confusión**

Confusion Matrix and ROC Curve

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

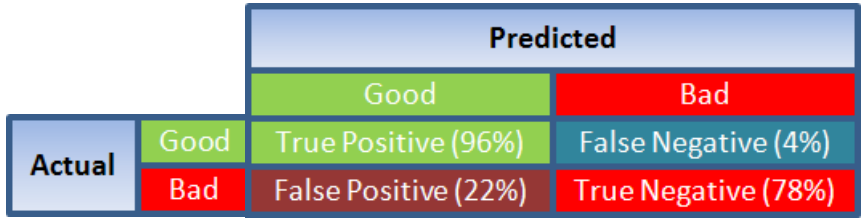
Model Performance

Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)

EVALUACIÓN DE MODELOS CREDIT SCORING

Existen diferentes indicadores para evaluar y comparar la capacidad predictiva de los modelos scorecard.

- **Curva ROC**
- **KS**
- **Gini**
- **Matriz de confusión**

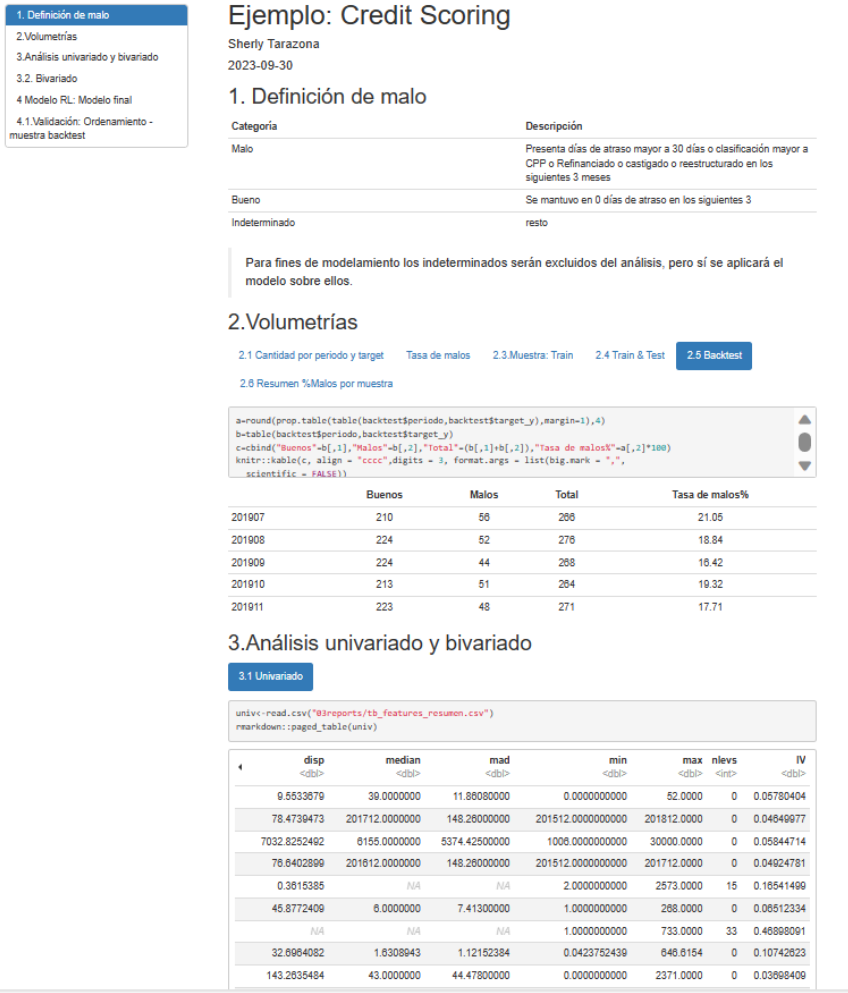


Parte práctica

Ejemplo:
Caso de aplicación: Modelo de Alertas tempranas.

Los datos han sido adecuados para efectos del curso.

Reporte con RMarkdown:



1. Definición de malo
2.Volumetrías
3.Análisis univariado y bivariado
3.2. Bivariado
4 Modelo RL: Modelo final
4.1.Validación: Ordenamiento - muestra backtest

Ejemplo: Credit Scoring

Sherly Tarazona

2023-09-30

1. Definición de malo

Categoría	Descripción
Malo	Presenta días de atraso mayor a 30 días o clasificación mayor a CPP o Refinanciado o castigado o reestructurado en los siguientes 3 meses
Bueno	Se mantuvo en 0 días de atraso en los siguientes 3
Indeterminado	resto

> Para fines de modelamiento los indeterminados serán excluidos del análisis, pero sí se aplicará el modelo sobre ellos.

2.Volumetrías

2.1 Cantidad por periodo y target | Tasa de malos | 2.3.Muestra: Train | 2.4 Train & Test | 2.5 Backtest | 2.6 Resumen %Malos por muestra

```
a=round(prop.table(table(backtest$periodo,backtest$target_y),margin=1),4)
b=table(backtest$periodo,backtest$target_y)
c=cbind("Buenos"=b[,1],"Malos"=b[,2],"Total"=(b[,1]+b[,2]),"Tasa de malos%"=a[,2]*100)
knitr::kable(c, align = "cccc",digits = 3, format.args = list(big.mark = ",",
  scientific = FALSE))
```

	Buenos	Malos	Total	Tasa de malos%
201907	210	56	266	21.05
201908	224	52	276	18.84
201909	224	44	268	16.42
201910	213	51	264	19.32
201911	223	48	271	17.71

3.Análisis univariado y bivariado

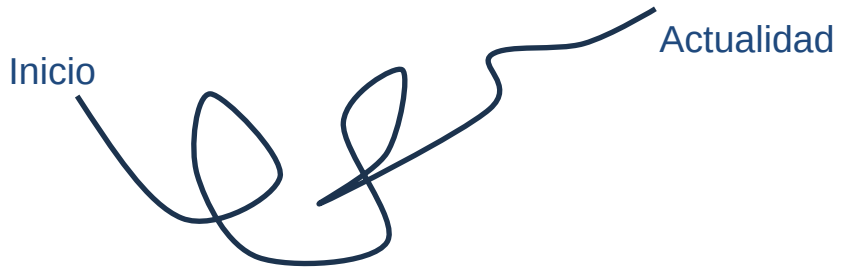
3.1 Univariado

```
univ<-read.csv("03reports/tb_features_resumen.csv")
rmarkdown::paged_table(univ)
```

disp <dbl>	median <dbl>	mad <dbl>	min <dbl>	max <dbl>	nlevs <int>	IV <dbl>
9.5533679	39.0000000	11.86080000	0.0000000000	52.0000	0	0.05780404
78.4739473	201712.0000000	148.26000000	201512.0000000000	201812.0000	0	0.04649977
7032.8252492	6155.0000000	5374.42500000	1006.0000000000	30000.0000	0	0.05844714
76.6402899	201612.0000000	148.26000000	201512.0000000000	201712.0000	0	0.04924781
0.3615385	NA	NA	2.0000000000	2573.0000	15	0.16541499
45.8772409	6.0000000	7.41300000	1.0000000000	268.0000	0	0.06512334
NA	NA	NA	1.0000000000	733.0000	33	0.46898091
32.6964082	1.6308943	1.12152384	0.0423752439	646.6154	0	0.10742823
143.2635484	43.0000000	44.47800000	0.0000000000	2371.0000	0	0.03698409

Recomendaciones

- Aprender a ser resilientes/ Reinversión constante
- La vida profesional no necesariamente va a ser recta
- Disfruten el proceso
- No tengas miedo de ser parte de grandes proyectos
- Que no te asuste el cambio y si te asusta, hazlo con miedo, pero hazlo.



Gracias

Sherly Tarazona Tocto

Advanced Analytics Lead |Co-organizer at R-Ladies Lima | Social Impact

linkedin.com/in/sherlytarazonatocto

Sherly Tarazona

https://linktr.ee/sherlytarazona

https://www.meetup.com/es/rladies-lima/