

Analítica Empresarial para la Toma de Decisiones Efectivas. ¿Cómo usar Big Data?

Rolando de la Cruz

Director Académico Magíster en Data Science

rolando.delacruz@uai.cl

Agenda

- ✓ Etapas de desarrollo de un producto de datos
- ✓ Algunas Soluciones Analíticas
- ✓ Aplicaciones
- ✓ Actividad

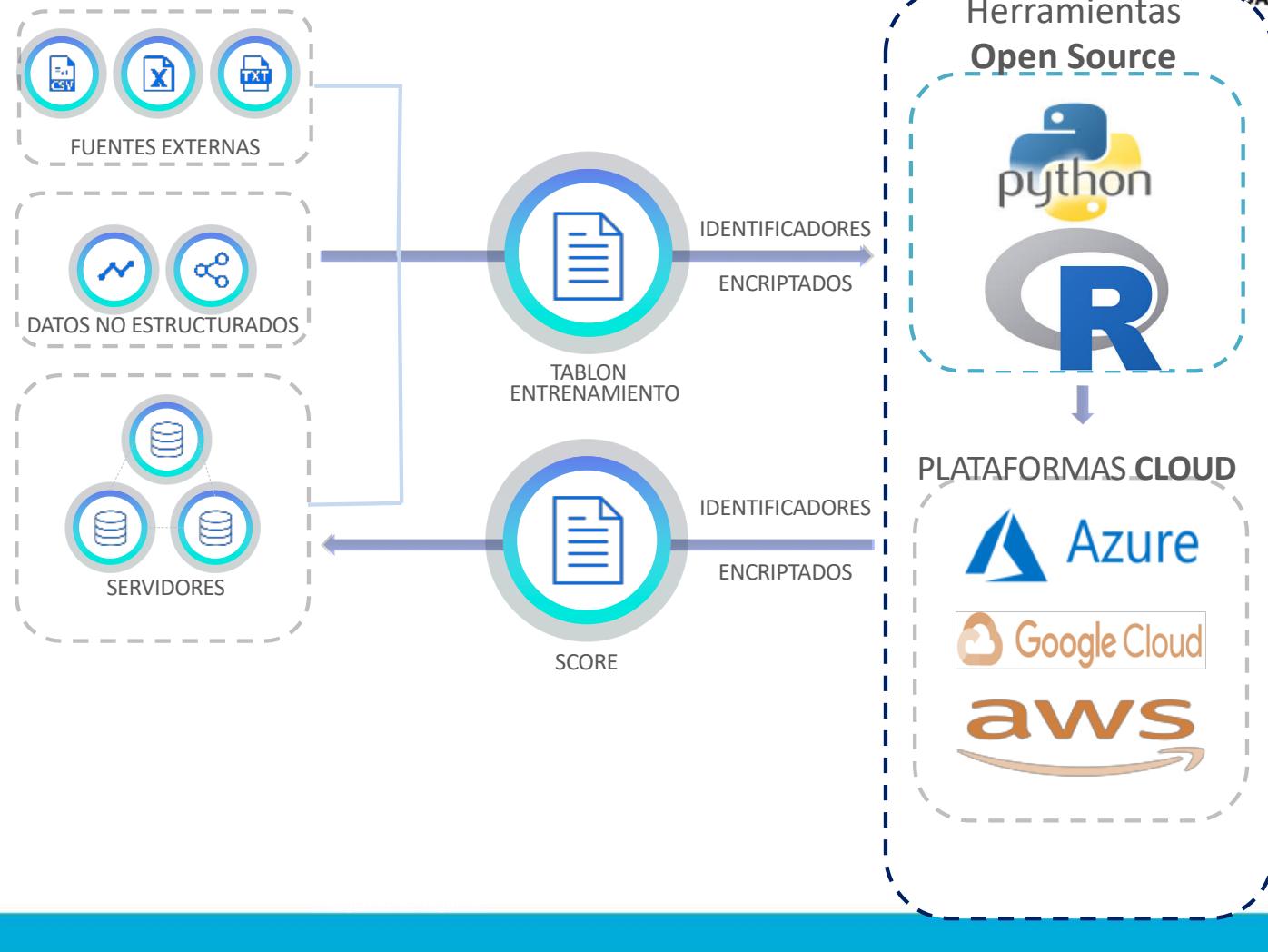
✓ Etapas de desarrollo de un producto de datos



FLUJO DE TRABAJO

- Las fuentes de información pueden estar en cualquier formato independiente del origen, todas ellas se concentrará en un tablón de datos para su posterior uso analítico.
- Las herramientas: R, Python, Ecosistemas Big Data, etc.
- Las plataformas Cloud son opcionales, en el caso que se decida considerar esta opción se puede utilizar para procesar el algoritmo seleccionado, expulsando el output a los servidores de la empresa
- Para dejar productivos los modelos y con regimen de procesamiento mensual (o el que sea definido) se debe evaluar la mejor opción que no implique ningún gasto adicional.

Flujo Trabajo

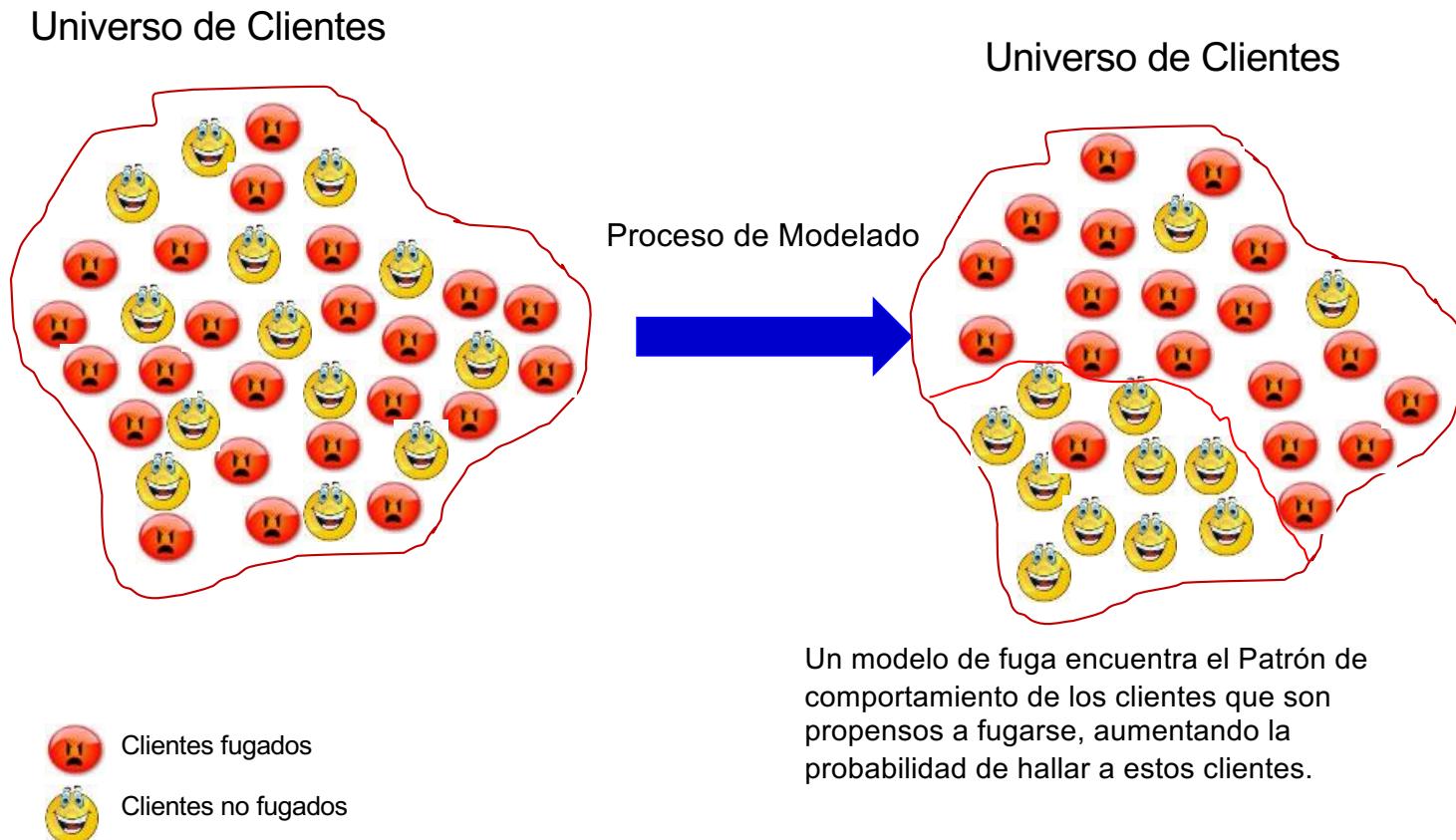


✓ Algunas Soluciones Analíticas

- a) Modelo de Fuga
- b) Modelo de Valor de Cliente
- c) Modelo de Segmentación

MODELO DE FUGA

MODELO DE FUGA



MODELO DE FUGA - Técnicas

Diversas técnicas estadísticas/machine learning para detectar fuga de clientes:

- Redes Neuronales - Anomaly Detection
- SVM - Regresión Logística
- Árboles de Decisión - Bosques Aleatorios, etc.

Estas técnicas extraen diferentes características (tanto categóricas como continuas) para así detectar posibles fugas.

El output es una etiqueta para cada “cliente”, donde la etiqueta de interés sería “ *posible fuga*”.

MODELO DE FUGA - Técnicas

- De las técnicas mencionadas anteriormente la **regresión logística** es la que entrega una interpretación del peso de cada una de las variables en la probabilidad de fuga, y conociendo esa relación se pueden tomar medidas para evitar la fuga.
- Otra ventaja de la regresión logística es que se puede calcular un scoring de fuga vía scorecard que son de fácil implementación en sistemas SQL.
- Con ello encontramos el patrón de comportamiento para predecir la *fuga*.

MODELO DE FUGA – ¿Cuál es el Patrón?

Penetración U6P	
Sin Dato	83
Menor al 0.001%	62
Entre 0.001% y 9.57 %	92
Entre 9.57% y 85.28 %	95
Mayor a 85.28%	96

Sexo	
Mujer	47
Hombre	98

Línea disponible u4p / RM	
Sin Dato	55
Menor a 0.016 veces rm	52
Entre 0.016 y 0.333 veces rm	90
Entre 0.333 y 2.5 veces rm	97
Entre 2.5 y 5.5 veces rm	93
Mayor a 5.5 veces rm	77

+

Nivel educacional	
No definido	68
Básico y Medio (completo e incompleto)	89
Técnico y universitario (completo e incompleto)	90
Técnico, técnico profesional y universitario	84

Cantidad de productos	
Menor a 2	80
Entre 2 y 3	86
Más de 3	91

Antigüedad del cliente	
Sin Dato	83
Menor a 8 meses	86
Entre 8 y 26 meses	91
Entre 26 y 62 meses	85
Entre 62 y 144 meses	78
Más de 144 meses	74

=

scorecard

MODELOS DE FUGA – ¿Cuál es el Output?

Los modelos entregan un score o ranking de clientes ordenados de mayor a menor(0 a 1000 puntos), donde 1000 es el puntaje con mayor probabilidad a realizar el evento estudiado (fuga).



RUT	SCORE
14.289.2XX-X	956,2
10.985.32X-X	841,9
9.548.62X-X	758,6
9.548.62X-X	625,3
5.868.61X-X	507,4
12.598.52X-X	468,5
7.948.23X-X	456,2
4.548.75X-X	389,1
8.635.47X-X	0,0

MODELO DE VALOR DE CLIENTE

VALOR DEL CLIENTE (Customer lifetime value - CLV)

Podemos saber en el pasado y hoy la rentabilidad de nuestros clientes, pero de cara al futuro debiésemos poder estimar cuánto tiempo más estarán con nosotros además de la rentabilidad que nos dejarán, estos dos pilares rentabilidad futura y ciclo de vida como cliente, nos aportan una medida que se denomina Valor del Cliente.

Valor de Cliente técnicamente se define como la suma de los márgenes de un cliente determinado desde el momento en que lo adquirimos hasta que deja de serlo.

VALOR DEL CLIENTE (Customer lifetime value - CLV)

Es la suma de los márgenes de un cliente determinado desde el momento en que lo adquirimos como cliente hasta que deja de serlo



Ejemplo: **-100 + 20 + 90 + 120 + 180 + + 175 = CLV**

¿Cuánto vale nuestro cliente hoy?

¿Cuál es el valor esperado de mi cliente de cara al futuro?

VALOR DEL CLIENTE (Customer lifetime value - CLV)

- El CLV es una métrica relacionada con el CRM (Customer Relationship Management).
- Es posible incluso confrontar la fuga de clientes versus otros indicadores como la rentabilidad por cliente, luego de una previa segmentación de la cartera de clientes, para así poder focalizar los esfuerzos en la retención de los clientes más apropiados, es decir, se podrá invertir los esfuerzos en retener aquellos clientes que le resultan más rentables.

¿Cómo lo hacemos? Fórmula

Valor de Cliente hoy = suma de los márgenes acumulados a la fecha

Valor futuro del Cliente

$$CFV_j = \sum_{t=1}^k m_{jt} * \frac{P_{jt}}{(1+d)^t}$$

- CFV_j = Valor futuro del cliente j (valor presente neto)
- t = periodo de tiempo
- k = tiempo proyectado
- P_{jt} = probabilidad de que el cliente j siga siendo nuestro en el periodo t
- m_{jt} = margen proyectado del cliente j en el periodo t
- d = tasa de descuento

MODELO CLV- Técnicas

Diversas técnicas estadísticas/machine learning para estimar el CLV. Van a depender si hay una relación contractual o no:

- Regresión lineal múltiple, Árboles de regresión, Bosques Aleatorios, etc.
- Modelo Pareto/NBD, Modelo de Cox. Modelo RFM (Recency, Frequency and Monetary value).

Estas técnicas extraen diferentes características (tanto categóricas como continuas) para así estimar el margen futuro y detectar posibles fugas.

Output: para cada “cliente” probabilidad de “*possible fuga*” y estimación del margen futuro.

¿Cuál es el output?

El modelo nos entrega por un lado la probabilidad de que un cliente permanezca ligado a la empresa en un cierto periodo, y por otro lado la proyección del margen en ese mismo periodo, y así podemos obtener el valor de cada cliente.

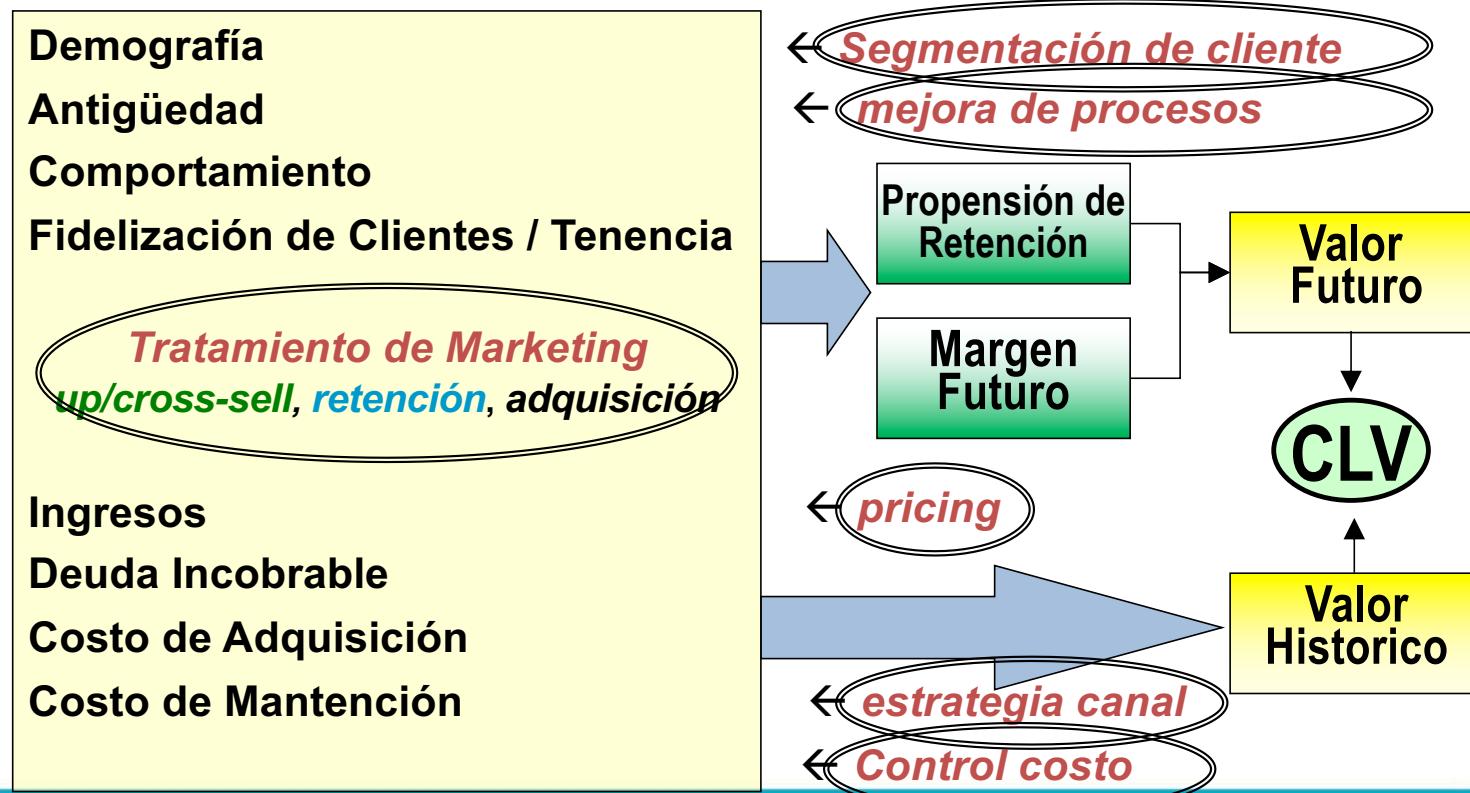
RUT	PERÍODO	PROBABILIDAD	MARGEN
12-k	1	0.99	150
12-k	2	0.95	152
12-k	3	0.90	145
12-k	4	0.83	146
12-k	5	0.75	135
12-k	6	0.71	135
12-k	7	0.62	120
12-k	8	0.58	120
12-k	9	0.53	119
12-k	10	0.51	110
12-k	11	0.48	105
12-k	12	0.45	105

Utilidades

- Poder proyectar las ganancias esperadas de toda la cartera de la empresa.
- Identificar o rankear a los clientes mediante su rentabilidad.
- Segmentar a los clientes.
- Tener una metrica sobre las acciones comerciales.

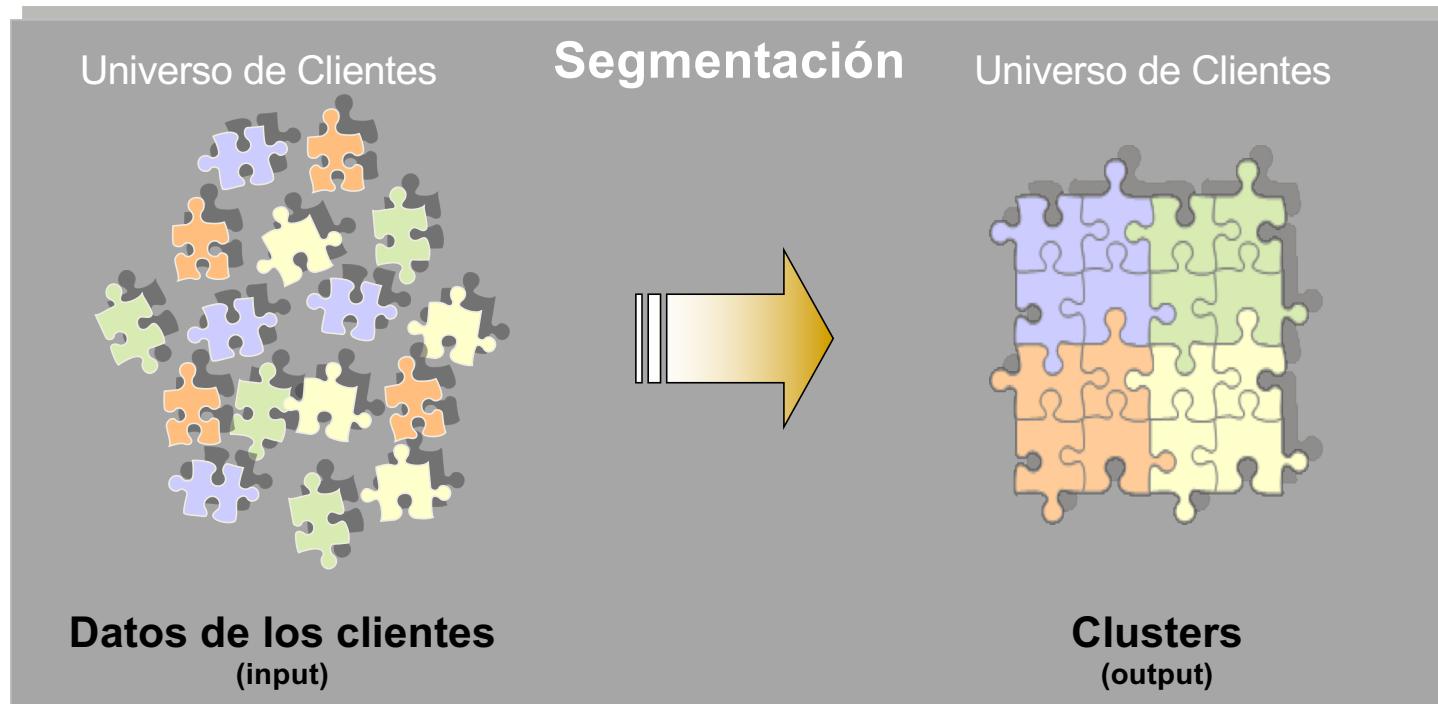
Mecanismo

...maximizar los retornos:



MODELO DE SEGMENTACIÓN

¿Qué es una Segmentación?



Ante la heterogeneidad de los clientes, una segmentación es la división del universo de clientes en grupos que tengan características similares, es decir, en grupos homogéneos para definir en función de estos últimos la oferta de servicios.

Segmentación

- **Segmentar** una cartera de clientes significa reconocer que está compuesto por diferentes individuos y que estos reaccionan de forma diferente a las propuestas de marketing.
- La institución reconoce que los productos y servicios que ofrece no sirven de la misma forma ni con la misma eficacia las expectativas de todos sus clientes.
- Para un acercamiento eficaz a los clientes la institución debe:
 - Definir grupos homogéneos de clientes: Grupos de clientes de acuerdo a características demográficas y otra información estadística.

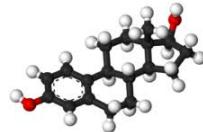
Segmentación

- Con una segmentación se puede determinar cuáles son los productos o servicios que se adaptan mejor a sus características y a sus necesidades económicas.
- Una vez determinados los distintos grupos se deben fijar los objetivos para el desarrollo del marketing. Por ejemplo, Campañas dirigidas.
- Usualmente la segmentación se realiza de acuerdo a objetivos propuestos.

✓ Aplicación

Predecir la evolución del embarazo

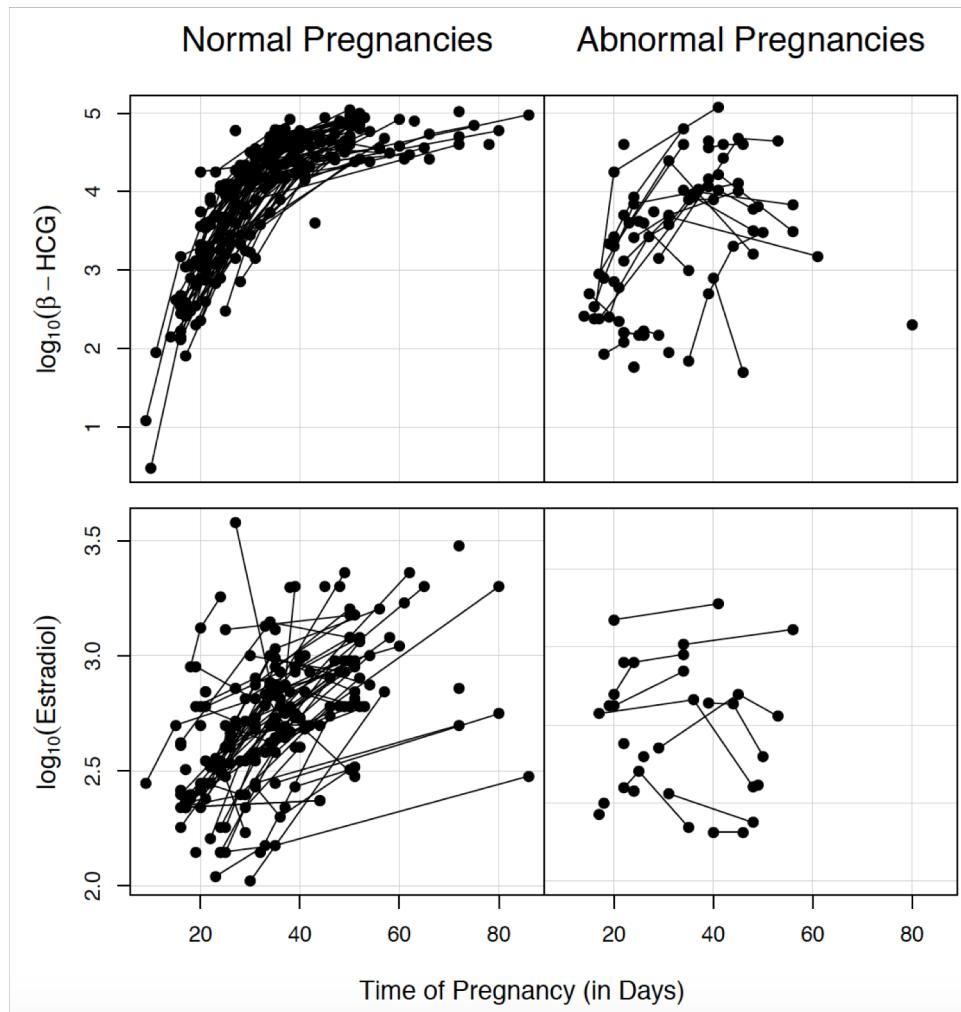
- ▶ En una clínica en Santiago, Chile se estudió el embarazo de 161 mujeres que se pueden clasificar en dos grupos:
 - ▶ Mujeres con embarazos normales ($n = 111$)
 - ▶ Mujeres con embarazos con pérdida espontánea del feto ($n = 50$)
- ▶ A cada mujer se le midió dos variables
 - ▶ Sub-unidad beta: hormona gonadotropina coriónica humana (HCG)
 - ▶ Estradiol



Publicación:

Marshall, G., De la Cruz-Mesía, R., Quintana, F. A., and Barón A. E. (2009). Discriminant analysis for multivariate longitudinal markers with possibly missing data. *Biometrics*, 65(1), 69-80.

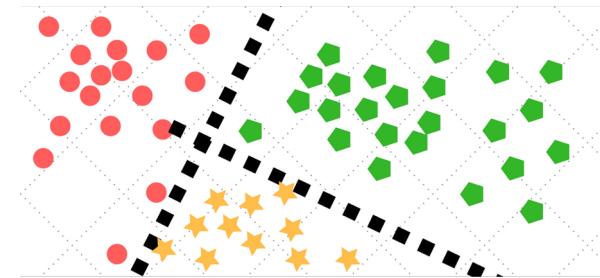
Perfiles Longitudinales



Subject i	Occasions t	Responses j			Covariates			
1	1	y_{111}	y_{121}	\cdots	y_{1r1}	x_{111}	\cdots	x_{1q1}
1	2	y_{112}	y_{122}	\cdots	y_{1r2}	x_{112}	\cdots	x_{1q2}
1	3	y_{113}	y_{123}	\cdots	y_{1r3}	x_{113}	\cdots	x_{1q3}
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\ddots	\vdots
1	s_1	y_{11s_1}	y_{12s_1}	\cdots	y_{1rs_1}	x_{11s_1}	\cdots	x_{1qs_1}
2	1	y_{211}	y_{221}	\cdots	y_{2r1}	x_{211}	\cdots	x_{2q1}
2	2	y_{212}	y_{222}	\cdots	y_{2r2}	x_{212}	\cdots	x_{2q2}
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\ddots	\vdots
2	s_2	y_{21s_2}	y_{22s_2}	\cdots	y_{2rs_2}	x_{21s_2}	\cdots	x_{2qs_2}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	1	y_{N11}	y_{N21}	\cdots	y_{Nr1}	x_{N11}	\cdots	x_{Nq1}
N	2	y_{N12}	y_{N22}	\cdots	y_{Nr2}	x_{N12}	\cdots	x_{Nq2}
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\ddots	\vdots
N	s_N	y_{N1s_N}	y_{N2s_N}	\cdots	y_{Nrs_N}	x_{N1s_N}	\cdots	x_{Nqs_N}

Clasificación

- ▶ Queremos clasificar individuos en uno de dos o más grupos.
- ▶ Las características de los individuos usada como información para la clasificación depende del tiempo y es medida longitudinalmente en intervalos arbitrarios y en número de ocasiones variable.
- ▶ Por la estructura no balanceada de los datos, no es posible usar técnicas clásicas de clasificación.



Trabajo previo + nuevos datos

- ▶ Marshall & Barón (2002) clasifican a mujeres embarazadas en parto normal o aborto espontáneo usando la subunidad-beta
- ▶ La subunidad-beta es un buen predictor del estado del embarazo en los primeros 80 días de edad gestacional
- ▶ Ahora incorporamos el nivel de Estradiol para complementar y mejorar la clasificación de las mujeres en estos dos grupos

Problema: Datos Faltantes

El porcentaje de datos faltantes en cada una de las variables y en cada uno de los grupos



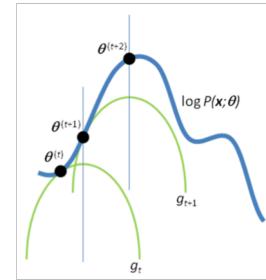
<http://www.jordicasanellas.com/data-science-blog/missing-data-impute-or-do-not-impute-r-examples>

Respuesta	Grupo	
	Normal	Con Pérdida
Sub-unidad beta	3	0
Estradiol	27	58

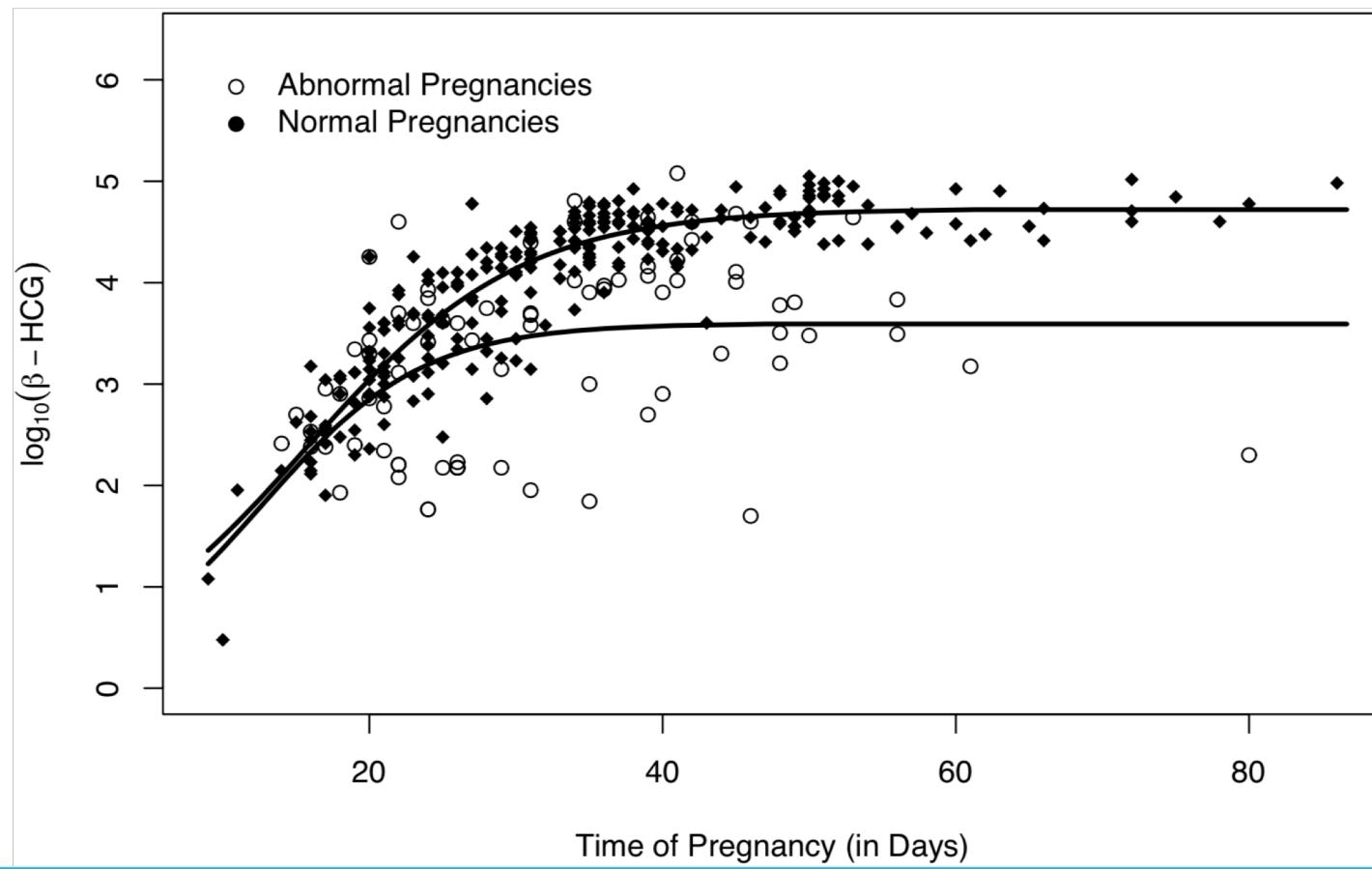
Subject i	Occasions t	Responses j				Covariates		
1	1	y_{111}	NA	...	y_{1r1}	x_{111}	...	x_{1q1}
1	2	y_{112}	y_{122}	...	y_{1r2}	x_{112}	...	x_{1q2}
1	3	y_{113}	y_{123}	...	NA	x_{113}	...	x_{1q3}
:	:	:	:	:	:	:	:	:
1	s_1	NA	NA	...	y_{1rs_1}	x_{11s_1}	...	x_{1qs_1}
2	1	NA	y_{221}	...	y_{2r1}	x_{211}	...	x_{2q1}
2	2	y_{212}	y_{222}	...	y_{2r2}	x_{212}	...	x_{2q2}
:	:	:	:	:	:	:	:	:
2	s_2	NA	y_{22s_2}	...	NA	x_{21s_2}	...	x_{2qs_2}
:	:	:	:	:	:	:	:	:
N	1	y_{N11}	y_{N21}	...	y_{Nr1}	x_{N11}	...	x_{Nq1}
N	2	y_{N12}	NA	...	NA	x_{N12}	...	x_{Nq2}
:	:	:	:	:	:	:	:	:
N	s_N	y_{N1s_N}	y_{N2s_N}	...	NA	x_{N1s_N}	...	x_{Nqs_N}

Desarrollo

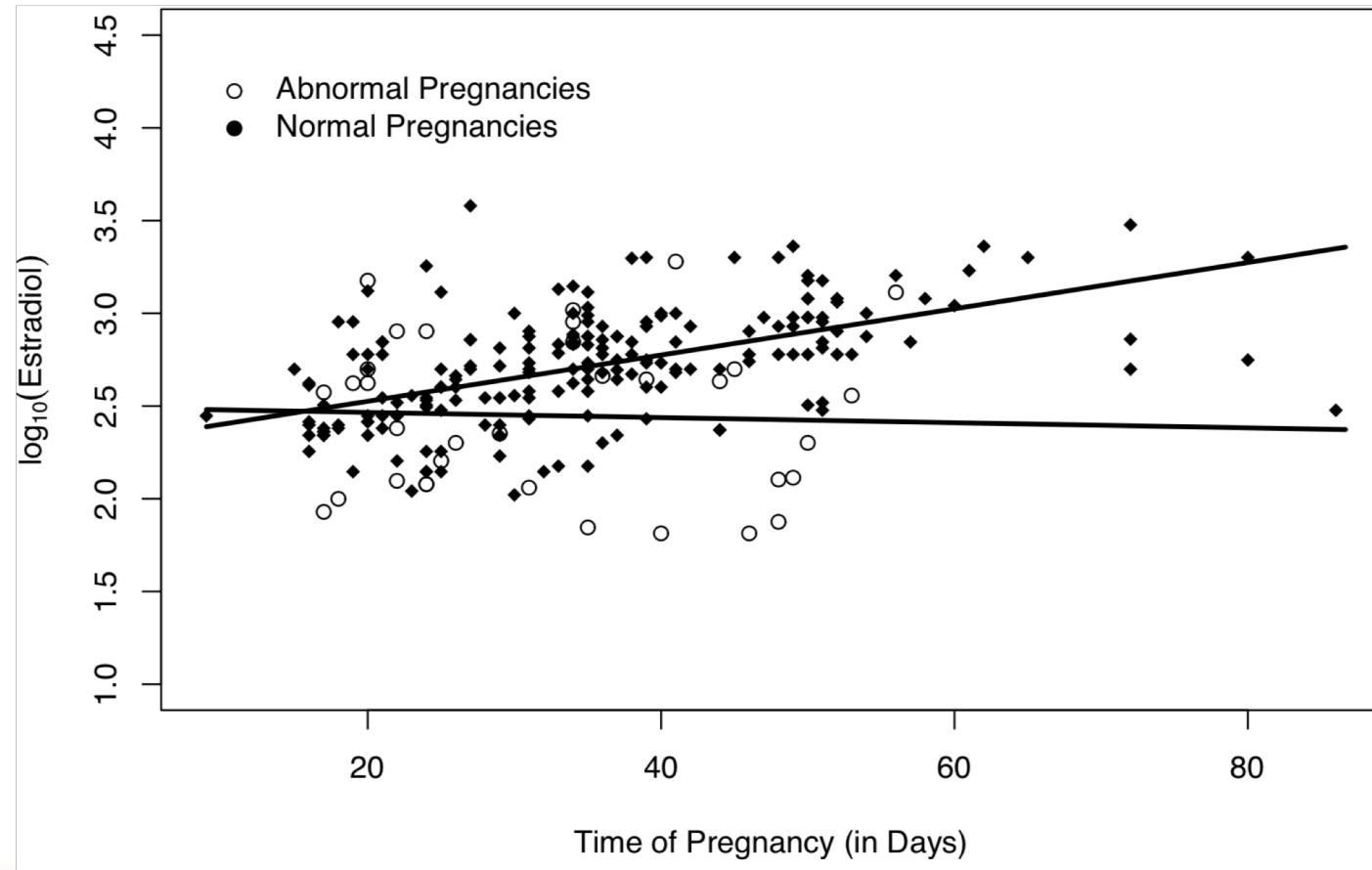
- Algoritmos de Estimación
- Implementación en plataforma open source

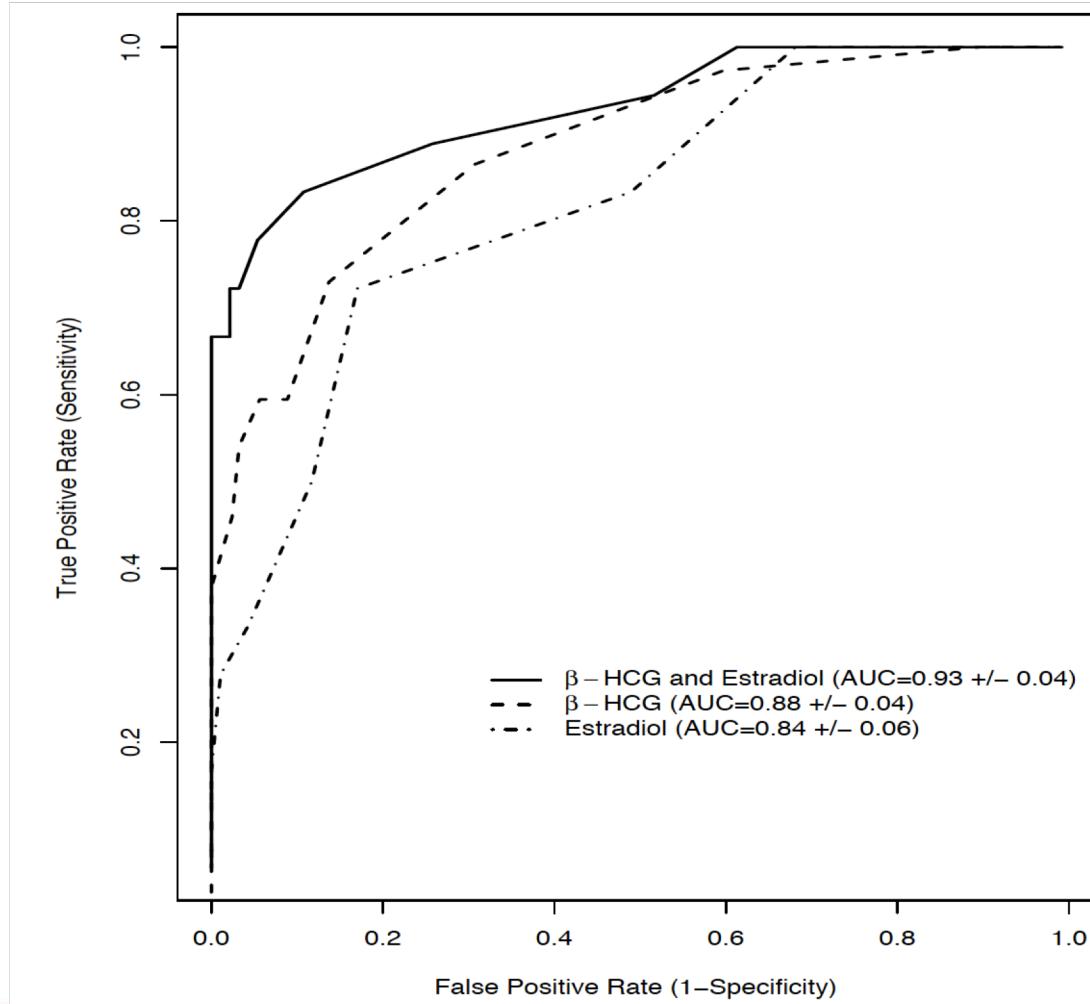


Modelos ajustados para Sub-unidad beta



Modelos ajustados para Estradiol





Curva ROC y AUC

Resultados

Classification results within-sample (A) and using Cross-validation (B)

Groups	Classification			
	(A)		(B)	
	Normal	Abnormal	Normal	Abnormal
Normal	117	7	117	7
Abnormal	16	21	17	20
	133	28	134	27
				161

Accuracy

- (A): 85,7%
- (B): 85,1%

Analítica Empresarial para la Toma de Decisiones Efectivas. ¿Cómo usar Big Data?

Rolando de la Cruz

Director Académico Magíster en Data Science

rolando.delacruz@uai.cl
