3TS: ML Engineering

TÉCNICAS BÁSICAS DE MODELADO PREDICTIVO

Objetivos

- Estructurar el proceso de desarrollo de un modelo
- Sistematizar las operaciones de exploración-preparación
- Entrenar modelos básicos
- Evaluar modelos

Objetivos

- Debatir, discutir y compartir experiencias y prácticas.
- Preguntarnos por qué

REPO: https://github.com/manualrg/DSLAB_Python

Índice de la sesión

- Vocabulario
- Visión holística del proceso de exploración-preparación
- Modelos básicos
- Evaluación de modelos (más allá de ROC)

Glossary

Features or predictors

Label, target or response

Prediction or scoring

id	x1	x2	…	y	$\hat{y}$_prob	$\hat{y}$_pred
cli101	1	1001		1 **event**	0.87	1
cli102				0	0.12	0

Example, instance or observation

Prior= AVG(y)

Posterior= AVG($\hat{y}$_prob)

Exploration-Feature Engineering

Numeric features

Low skewness

High skewness

OI_: Outlier idx
MI_: NaN idx
Missing imp.

Binning
Bucketing
Transformation
Normalization [0,1]
Standartization $\{\mu = 0,\ \sigma = 1\}$

Categorical features

Low cardinality

High cardinality

Missing imp.
Rare levels
Min cell freq

Low freq grouping
Numeric mapping
OHE
Dense representation
Embeddings

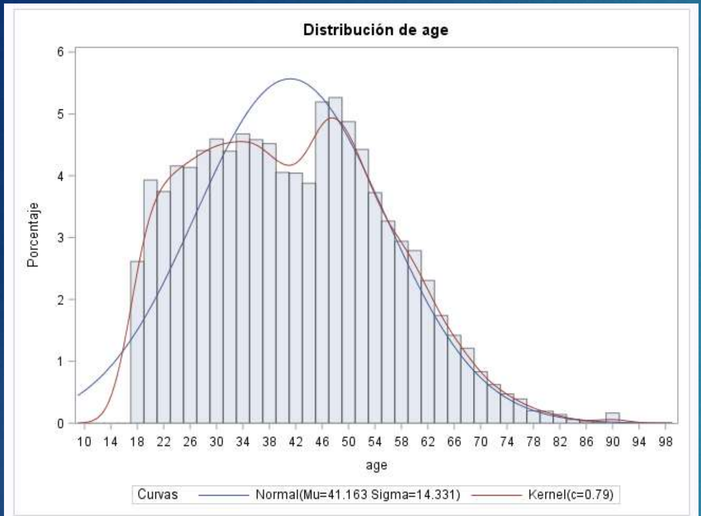
Metadata Analysis

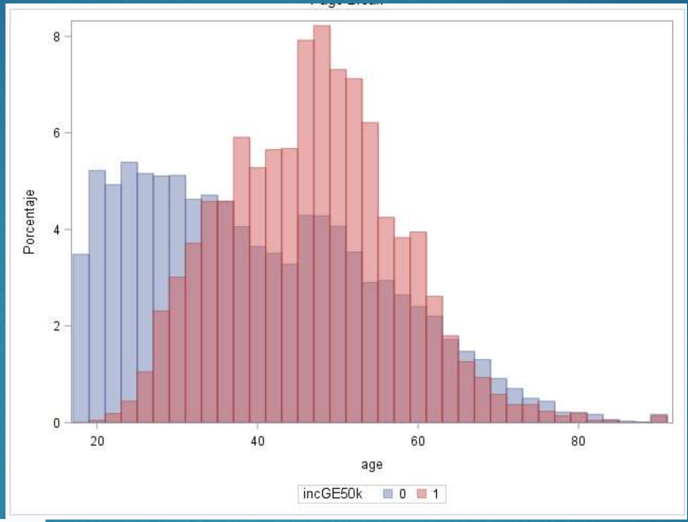
Descriptive Stats
checkMissing()
checkSkewness()
checkCatFreq()

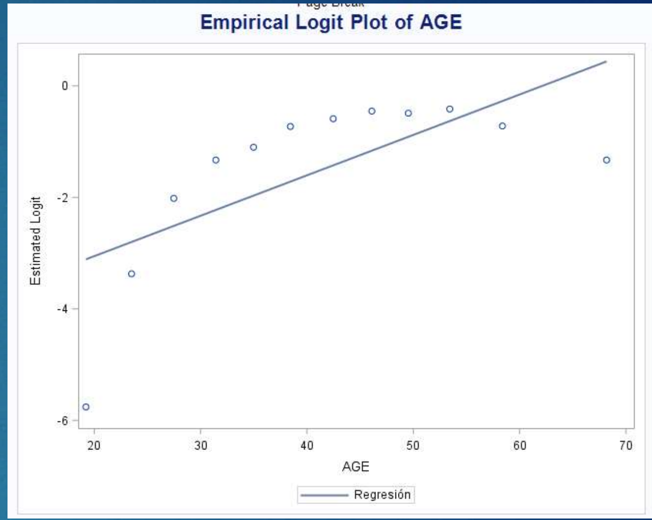
Variable Screening
screenMissing()
screenOutliers()
screenLowFreq()

Feature Engineering

Numeric Features: Low Skewness







raw numeric feature

Procedimiento HPLOGISTIC

Estadísticas de ajuste de partición		
Estadístico	Entrenamiento	Validación
Área bajo ROCC	0.6522	0.6499

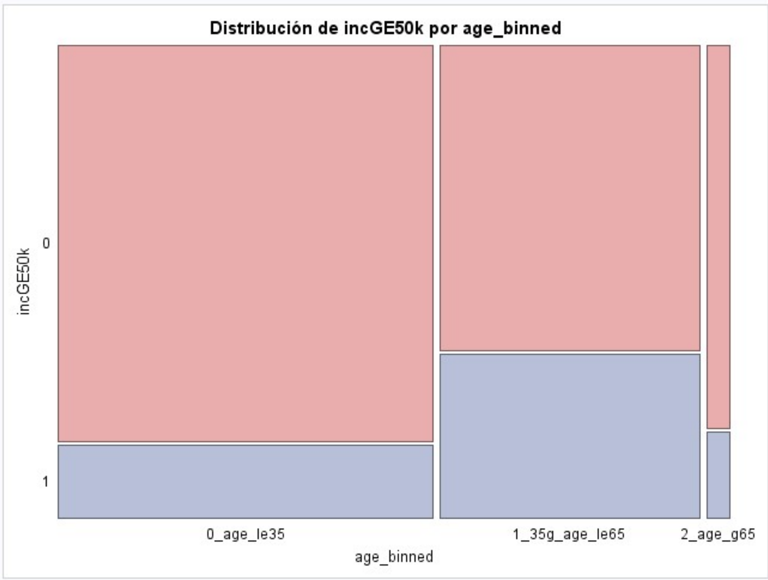
transformed numeric feature

Procedimiento HPLOGISTIC

Estadísticas de ajuste de partición		
Estadístico	Entrenamiento	Validación
Área bajo ROCC	0.6945	0.6934

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Numeric Features: Low Skewness



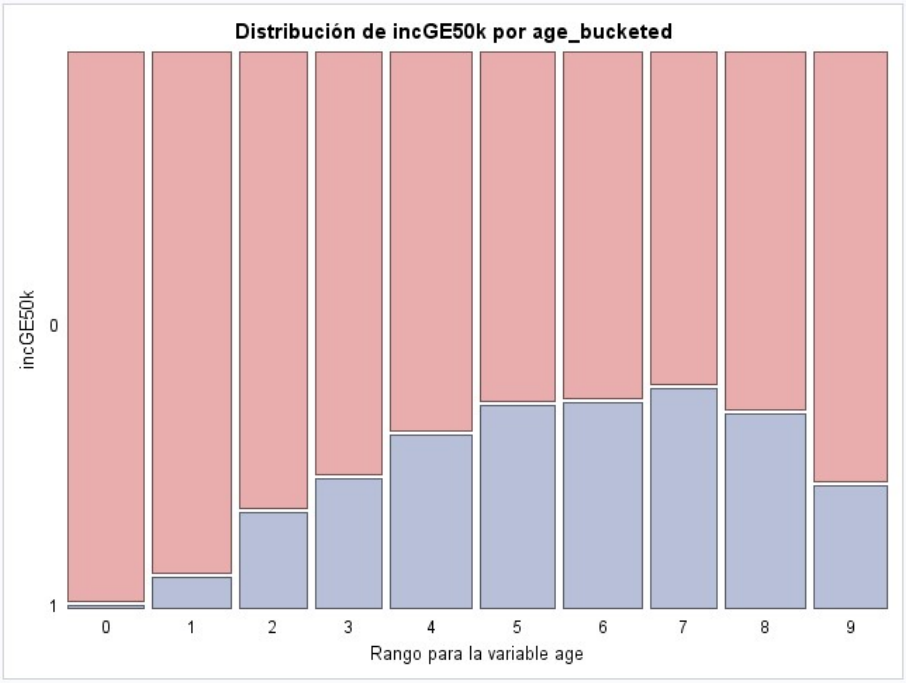
Variable de análisis : age						
age_binned	Número de observaciones	N	Media	Dev std	Mínimo	Máximo
0_age_le35	13927	6679	26.6548885	5.2977745	17.0000000	35.0000000
1_35g_age_le65	9640	9640	48.5193983	7.8466232	36.0000000	65.0000000
2_age_g65	853	853	71.6213365	5.6303057	66.0000000	90.0000000

binned numeric feature

Procedimiento HPLOGISTIC

Estadísticas de ajuste de partición		
Estadístico	Entrenamiento	Validación
Área bajo ROCC	0.6304	0.6297

Numeric Features: Low Skewness



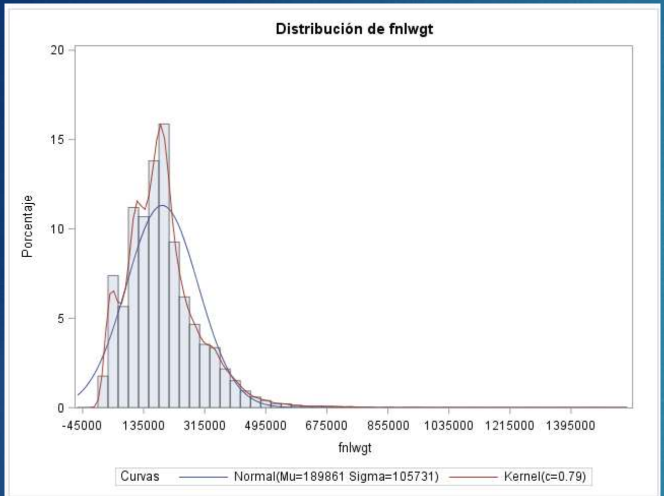
Variable de análisis : age						
Rango para la variable age	Número de observaciones	N	Media	Dev std	Mínimo	Máximo
0	1767	1767	19.7475948	1.6127724	17.0000000	22.0000000
1	1812	1812	25.0336645	1.4314729	23.0000000	27.0000000
2	1558	1558	29.5577664	1.1111339	28.0000000	31.0000000
3	1542	1542	33.5421530	1.0988638	32.0000000	35.0000000
4	1875	1875	37.9189333	1.4182975	36.0000000	40.0000000
5	1728	1728	43.0104167	1.4247812	41.0000000	45.0000000
6	1830	1830	47.3726776	1.1164641	46.0000000	49.0000000
7	1531	1531	51.3742652	1.1145344	50.0000000	53.0000000
8	1849	1849	56.8253110	2.0000778	54.0000000	60.0000000
9	1680	1680	67.2523810	6.0640315	61.0000000	90.0000000

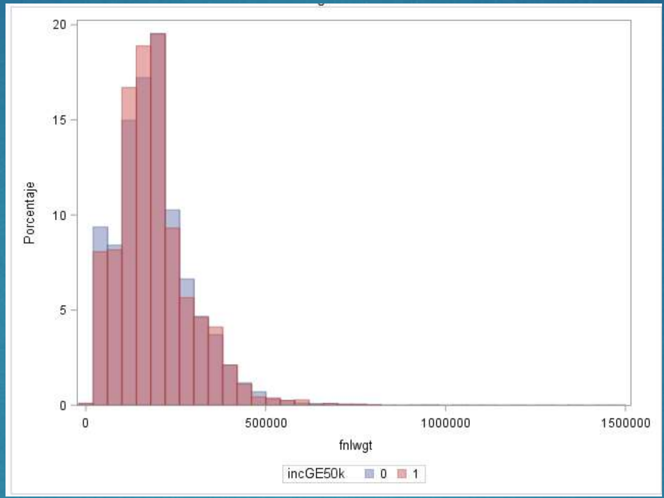
bucketed numeric feature

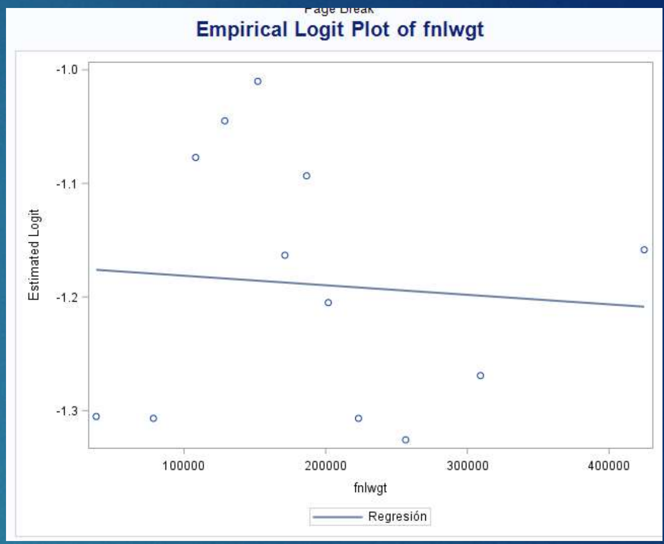
Procedimiento HPLOGISTIC

Estadísticas de ajuste de partición		
Estadístico	Entrenamiento	Validación
Área bajo ROCC	0.6930	0.6915

Numeric Features: High Skewness







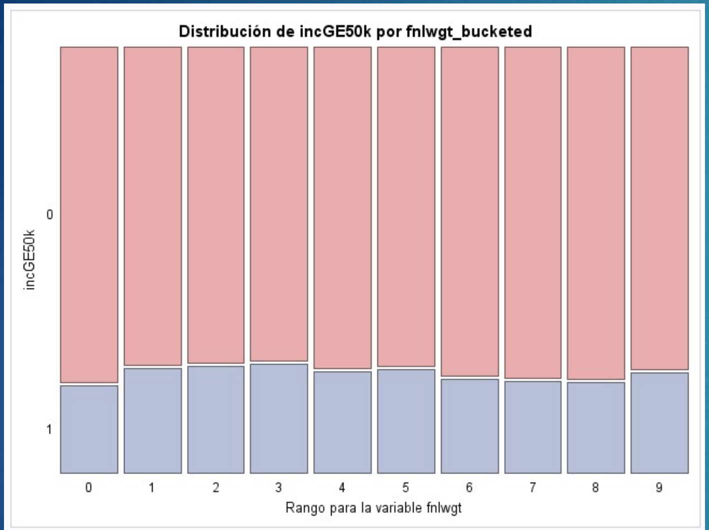
Page Break

raw numeric feature

Procedimiento HPLOGISTIC

Estadísticas de ajuste de partición		
Estadístico	**Entrenamiento**	**Validación**
Área bajo ROCC	0.5055	0.5067

Numeric Features: High Skewness



Variable de análisis : fnlwgt						
Rango para la variable fnlwgt	Número de observaciones	N	Media	Dev std	Mínimo	Máximo
0	2441	2441	41663.27	11901.64	12285.00	65368.00
1	2443	2443	89261.97	12006.26	65372.00	106437.00
2	2442	2442	118040.10	6782.76	106491.00	130557.00
3	2442	2442	145338.68	7854.66	130571.00	158712.00
4	2441	2441	169418.24	5866.00	158734.00	178778.00
5	2443	2443	187964.40	4987.43	178780.00	196791.00
6	2443	2443	207311.38	6555.10	196797.00	219838.00
7	2442	2442	238208.39	11349.59	219841.00	259496.00
8	2441	2441	291215.68	19980.83	259505.00	329759.00
9	2442	2442	410198.32	98322.91	329783.00	1484705.00

bucketed numeric feature

Procedimiento HPLOGISTIC

Estadísticas de ajuste de partición		
Estadístico	Entrenamiento	Validación
Área bajo ROCC	0.5269	0.5254

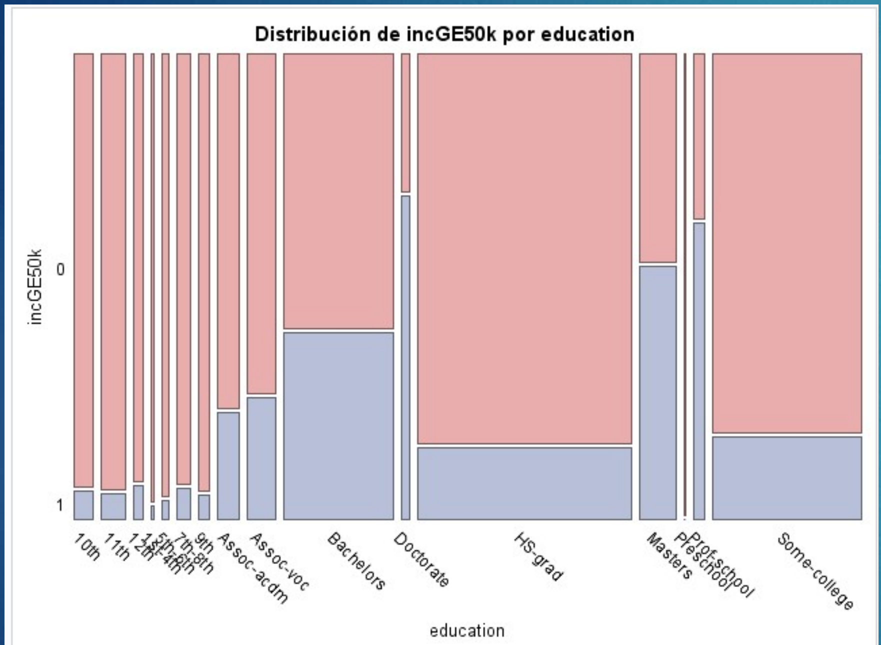
Categorial Features

- Nominal levels (as strings): OHE
- Ordinal levels (as numeric)
- Numeric mapping:
  - Freq count
  - Freq idx
  - Event proportion
- …

Design matrix (Sparse representation)

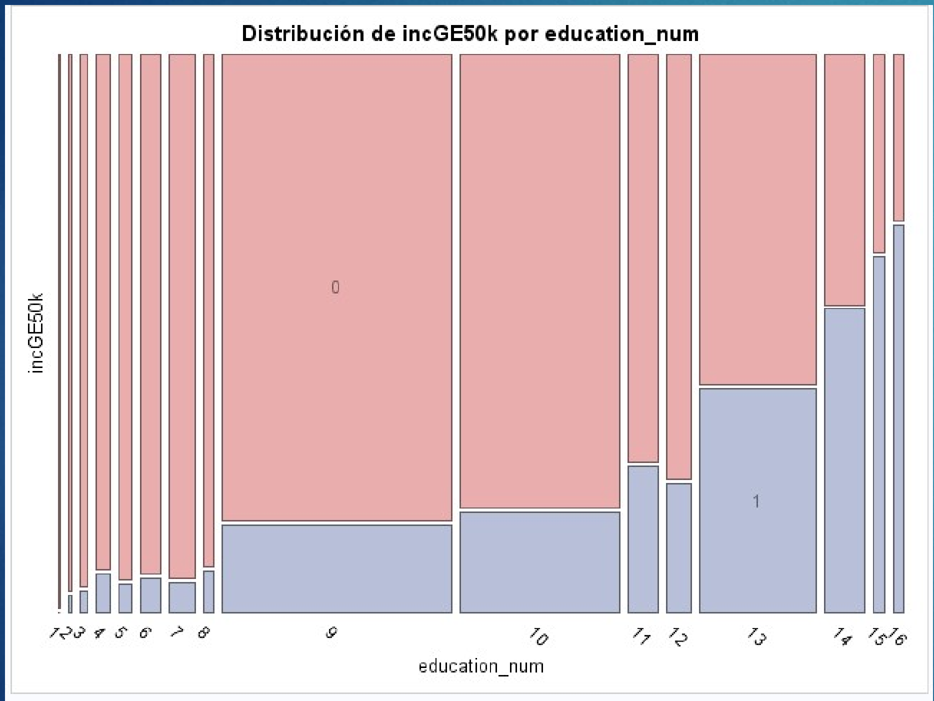
Información de nivel de clase																	
Clase	Valor	Diseño de variables															
education	10th	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	11th	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	12th	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	1st-4th	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	5th-6th	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	7th-8th	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	9th	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	Assoc-acdm	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	Assoc-voc	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	Bachelors	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	Doctorate	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	HS-grad	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	Masters	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	Prof-school	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	Some-college	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	Preschool	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Categorial Features: Nominal



Estadísticas de ajuste de partición		
Estadístico	**Entrenamiento**	**Validación**
Área bajo ROCC	0.7162	0.7025
[illegible]	[illegible]	[illegible]

Categorial Features: Ordinal



Page Break

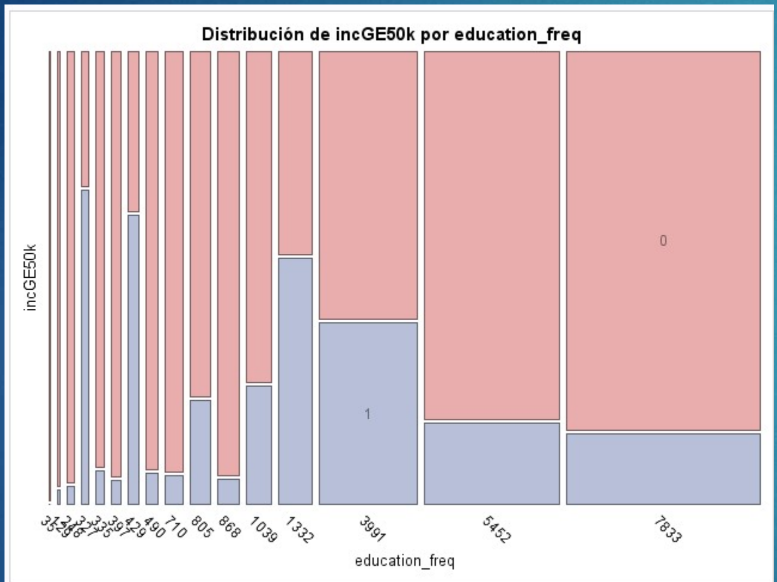
Ordinal: education_NUM

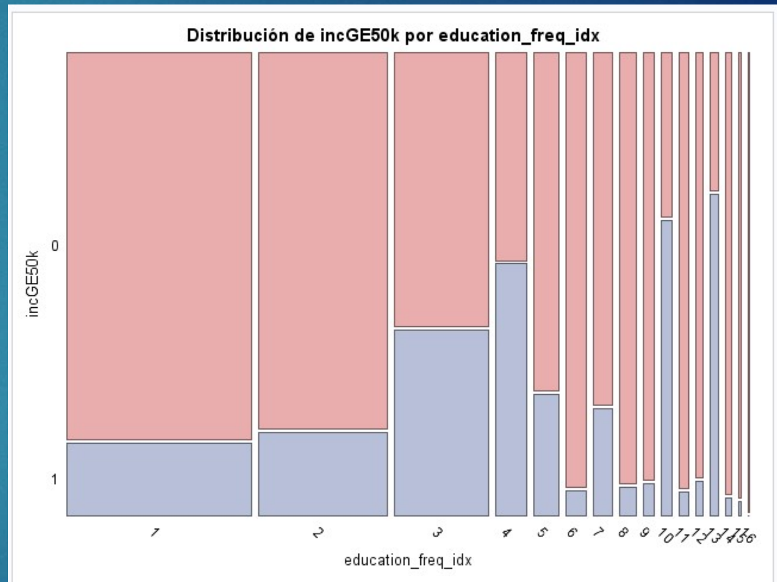
Procedimiento HPLOGISTIC

Estadísticas de ajuste de partición		
Estadístico	**Entrenamiento**	**Validación**
Área bajo ROCC	0.7162	0.7025

Magical numbers?
Distances?

Categorial Features: Freq mappings





Estadísticas de ajuste de partición		
Estadístico	Entrenamiento	Validación
Área bajo ROCC	0.5791	0.5709

Estadísticas de ajuste de partición		
Estadístico	Entrenamiento	Validación
Área bajo ROCC	0.5795	0.5712

Basic Modelling: Logistic Regression

4.3.4 Multiple Logistic Regression

We now consider the problem of predicting a binary response using multiple predictors. By analogy with the extension from simple to multiple linear regression in Chapter 3, we can generalize (4.4) as follows:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{4.6}$$

where $X = (X_1, \ldots, X_p)$ are p predictors. Equation 4.6 can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \tag{4.7}$$

Just as in Section 4.3.2, we use the maximum likelihood method to estimate $\beta_0, \beta_1, \ldots, \beta_p$.

Logistic Regression Model

Want $\boxed{0 \le h_\theta(x) \le 1}$

$$h_\theta(x) = g(\theta^T x)$$

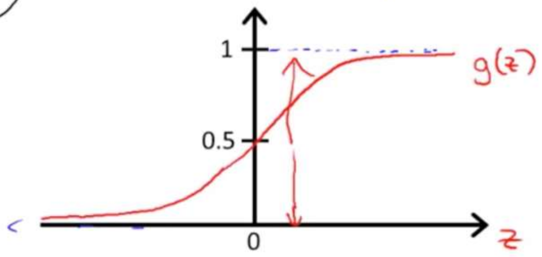
$$\rightarrow g(z) = \frac{1}{1+e^{-z}}$$

$\theta^T x$

$$\boxed{h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}}$$

→ Sigmoid function

→ Logistic function



Parameters θ.

Gradient Descent

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log\left(1-h_\theta(x^{(i)})\right)\right]$$

Want $\min_\theta J(\theta)$:

Repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

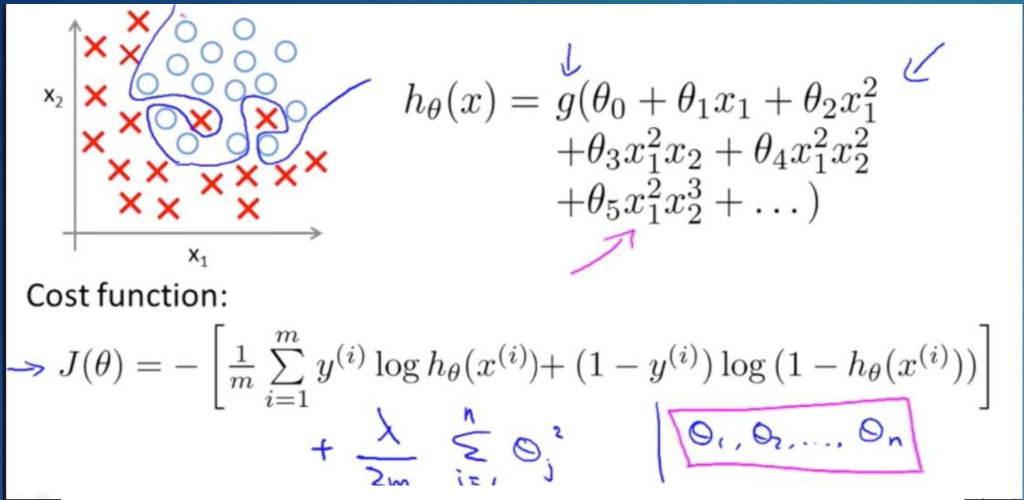
}

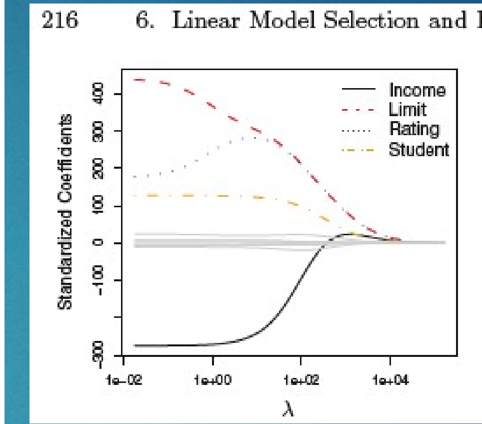
(simultaneously update all θ_j)

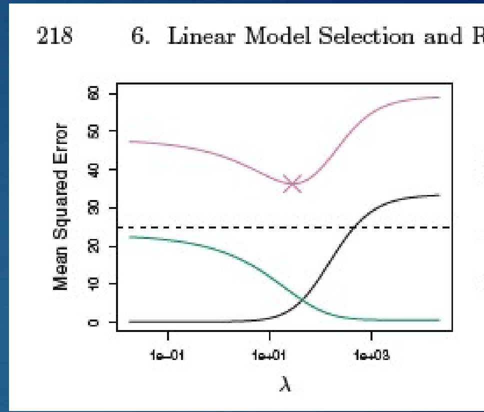
$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

Andrew Ng

Basic Modelling: Regularization







228 6. Linear Model Selection and Regularization

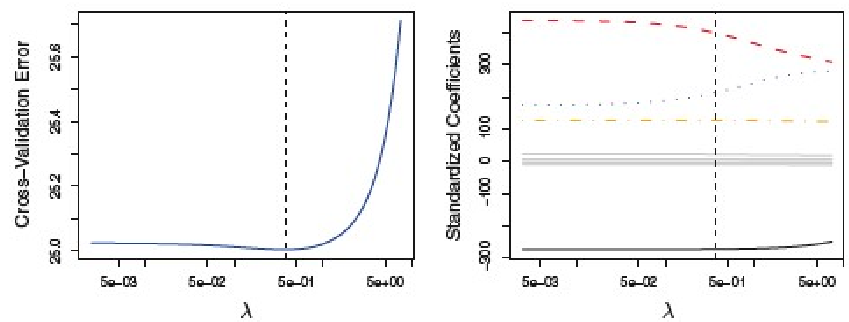
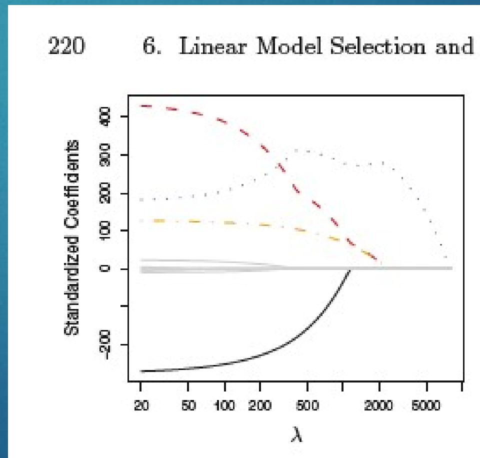
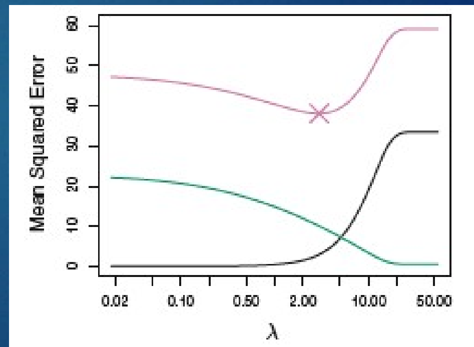
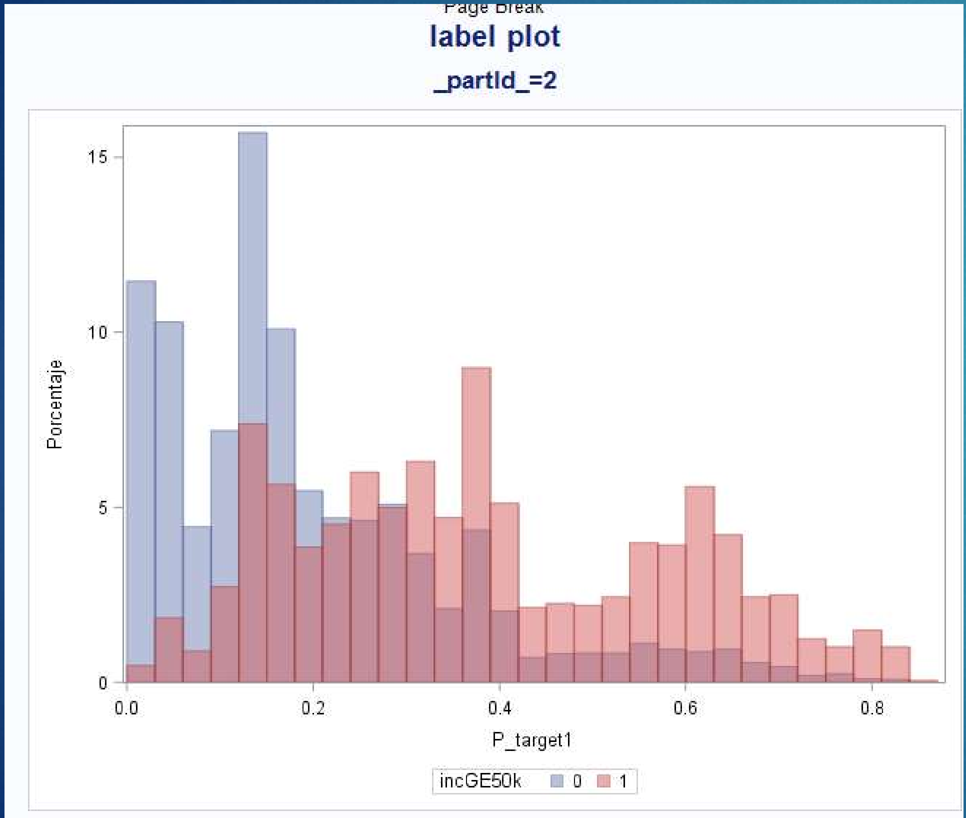


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* Credit *data set with various value of λ.* Right: *The coefficient estimates as a function of λ. The vertical dashed lines indicate the value of λ selected by cross-validation.*



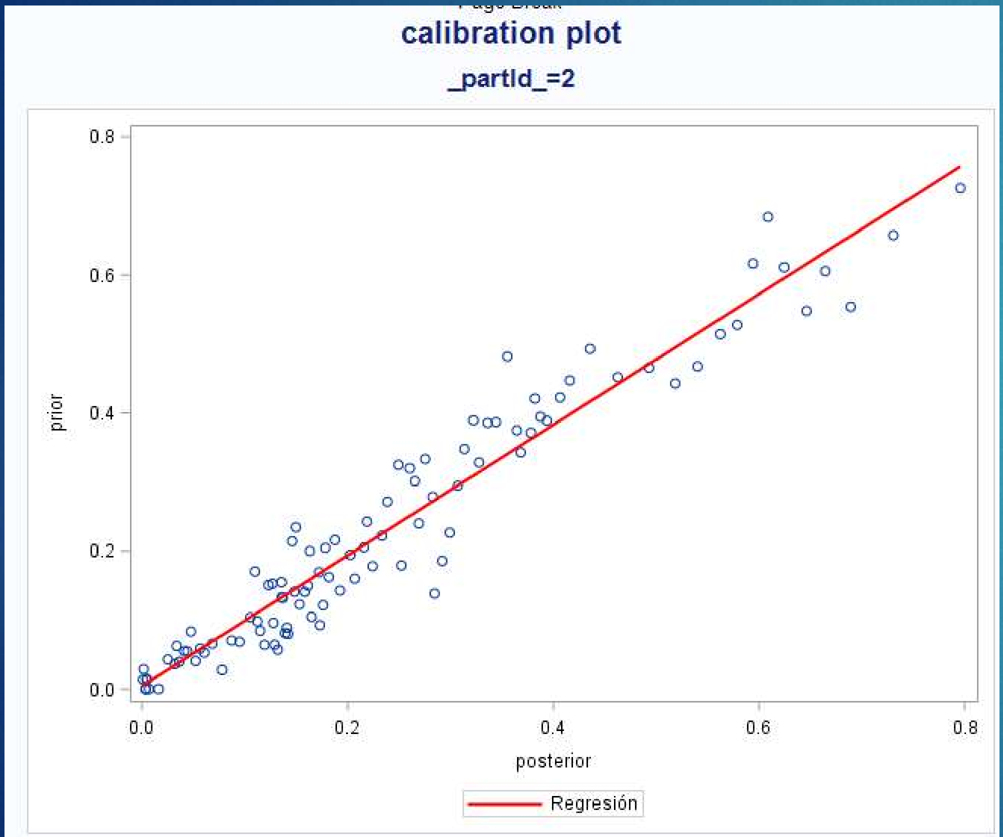


Model Assessment: Beyond ROC



- Models that classify low probability examples correctly as non-event can yield a high AUC (>0.99) and not perform properly
- RARE EVENT BINARY CLASSIFICATION
- What is AUROC?

Model Assessment: Beyond ROC



- Class separation (KS metric)
- PR Curves
- Analyze scoring by predicted probability bucket!!! (e.g. calibration plot, Lift, Gain.)

GRACIAS !!!

- BIBLIOGRAFÍA:
- Introduction to Statistical Learning (R)
  - https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
- Elements of Statistical Learning (R)
  - https://web.stanford.edu/~hastie/Papers/ESLII.pdf
- Machine Learning Andrew NG (Matlab-Octave)
  - https://www.coursera.org/learn/machine-learning
- Categorical Data Analysis Using Logistic Regression (SAS)
  - https://support.sas.com/edu/schedules.html?ctry=us&crs=CDALR
- Predictive Modeling Using Logistic Regression (SAS)
  - https://support.sas.com/edu/schedules.html?ctry=us&crs=PMLR