Agenda



- anuncios varios
 - Tarea 3 entrega lunes 8 de Junio (Preguntas)
- modelos de analitica (machine learning-ML) Supervisado
 - Redes Neuronales
 - LSTM
 - Secuencias
- Evaluación de modelos
 - Métricas
 - Monitoreo de modelos
 - Estabilidad
 - Validez
- Spark

"workflow" Flujos de trabajo para crear un modelo



DATA TRANSFORMATION ————————————————————————————————————			MODELOS ANALITICA		
leer datos (data pulling)	explorar variables	transformar datos	sampling split (muestra)	entrenar modelo	evaluar modelo
leer las fuentes de ⁻ datos y explorar - -	calidad en los - datos, nulos - distribuciones - encontrar outliers	imputar - centrar, escalar agrupar variables con - muchas categorías	split por tiempo, % de datos. subsampling, oversampling	 utilizar datos de training y evaluar en test 	evaluar los resultado del modelo en múltiples segmentos



Accuracy:

• Clasificaciones correctas/ Total clasificaciones

Labels \ Predicción	No Paso	Paso
No Paso	4 (True Negative)	1 (False Positive)10
Paso	1 (False Negative)	14 (True Positive)

accuracy =
$$(TN + TP)/(TN+FP+FN+TP) = 18/20=0.9$$

precisión = $TP/(TP+FP)=14/15$ (cols)=0.933 (false positive)
recall = $TP/(FN + TP)=14/15$ (rows)=0.933 (false negative) hit rate

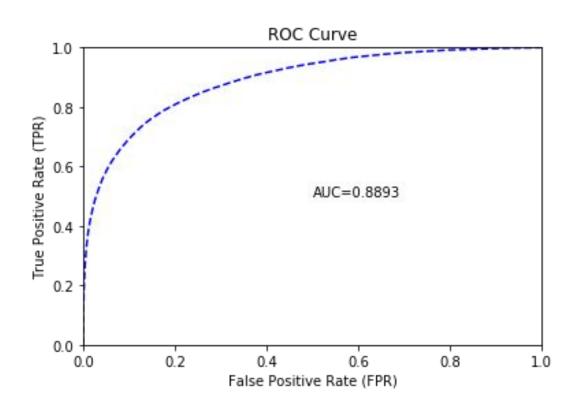
Fl score:

Es la media armónica entre precision y recall = 2 (precision* recall) / (precision + recall)

$$F1 = 2 TP/(2TP + FP + FN)$$



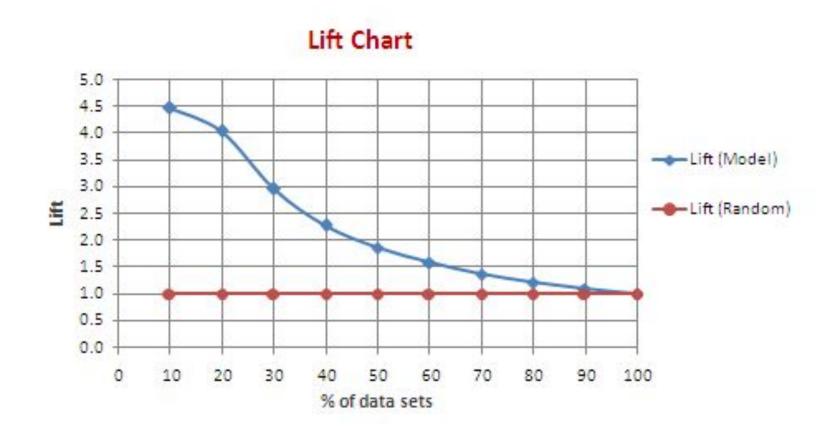
AUC Área bajo la curva: Es especial para datos desbalanceados



	AUC	% fraude en 4 Millones
GBM (Arboles)	0.88	0.71



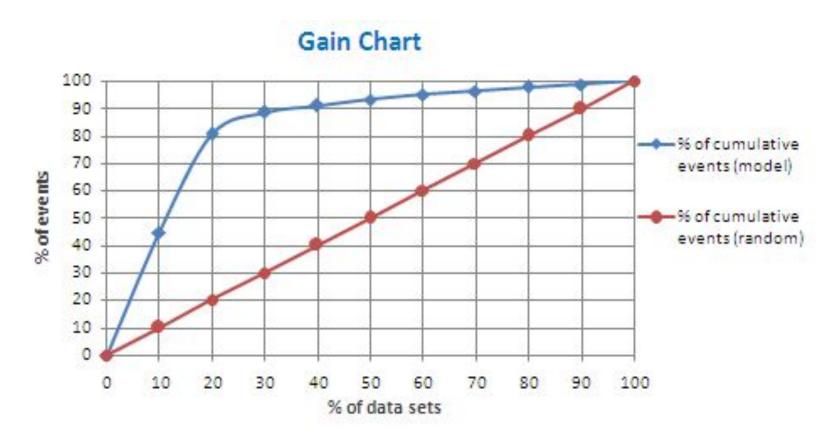
Lift chart: hit rate, capture o porcentaje de datos que se predicen bien





Lift chart acumulado o gain chart:

• valor acumulado de hit rate, capture o porcentaje de datos que se predicen bien

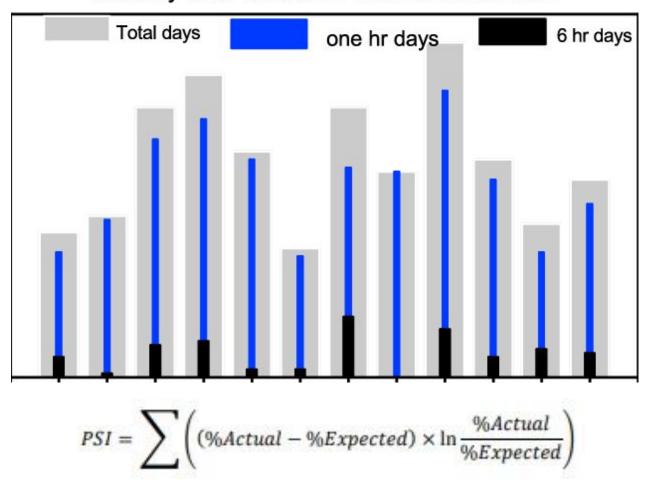


Métricas de estabilidad



Histogram por mes:

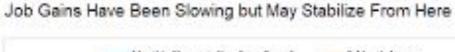
Monthly data from Jan. 1998 to Dec. 2009



Métricas de validez



Lift chart por mes:





Source LPL Research, U.S. Bureau of Labor Statistics 10702