Welcome

Data Mining Economía y Finanzas - 2022 - Comisión 1

Alejandro Bolaños + Ivo Rusconi

Sobre Mí, Sobre Ustedes

- ¿Quiénes somos?
- Check in: de que se trata.
- Sobre sus cursadas del cuatrimestre anterior

Sobre esta Materia > Objetivo

Resolver un problema de dimensiones reales del mercado argentino utilizando las herramientas tecnológicas para manejar grandes volúmenes de datos y ser capaz de generar una predicción competitiva en el mercado profesional laboral argentino.

Sobre esta Materia > ¿Qué vamos a ver?

- Metodologías

  Aprender a pensar

- Algoritmos y técnicas de última generación de modelado predictivo sobre datos estructurados.

  Usar las mejores herramientas

- MLOps

  Saltar de las notebook a los procesos

- Big Data en la nube

  Vamos a jugar a lo grande

Sobre esta Materia > ¿Cómo lo vamos a ver?

Sincrónicamente en Clases presenciales (Lunes + Plenarias)

- Escucharnos
- Actividades - (traigan Notebook cada 2 o 3 alumnos)
- Presentaciones de compañeros

Asincrónicamente a través de Zulip

- Consultas
- Desarrollo de experimentos
- Memes

Sobre esta Materia > ¿Cómo lo vamos a ver?

4 Competencias escalonadas (mismo problema)

% 5 + 10 + 10 + 25

2 videos

% 9 + 9

23 Experimentos

% 20

1 examen

% 1

Otras zulip (10%) + cazatalentos (+1p xor +2p)

Sobre esta Materia > 23 EXPERIMENTOS?!

No todos van a hacen todos. Son experimentos de construcción colectiva.

El fin es ayudarles a experimentar de forma correcta y que los resultados ayuden también a sus compañeros a entregar mejores modelos.

Sobre esta Materia > Qué vamos a usar?

XGBoost

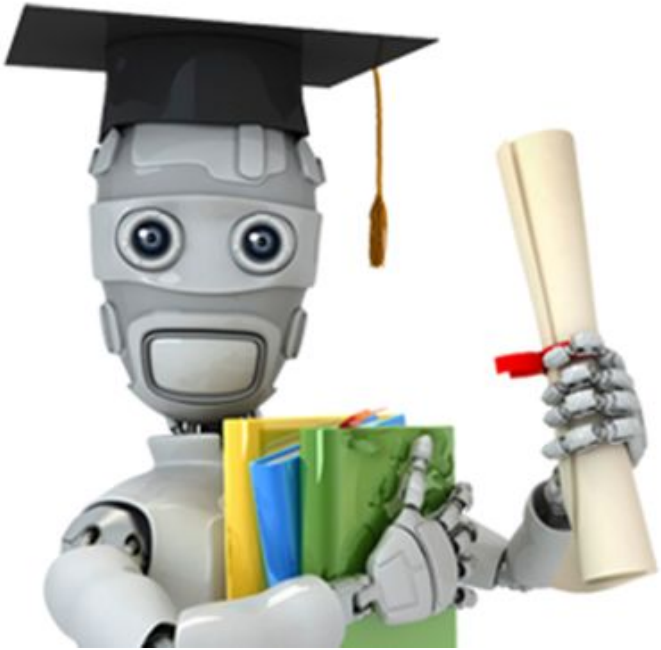
data.table

MLRmbo

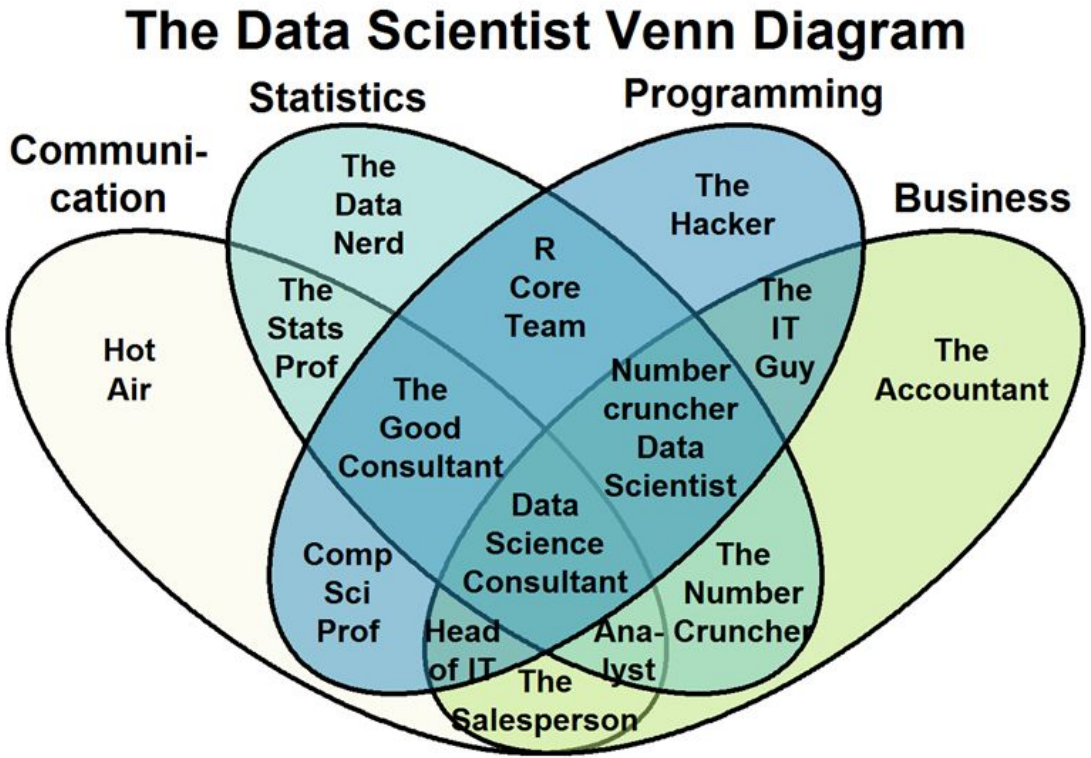
mlflow™

Pero NO hay limitaciones, usted puede usar lo que quiera

Storytelling



Storytelling > Usted, un unicornio



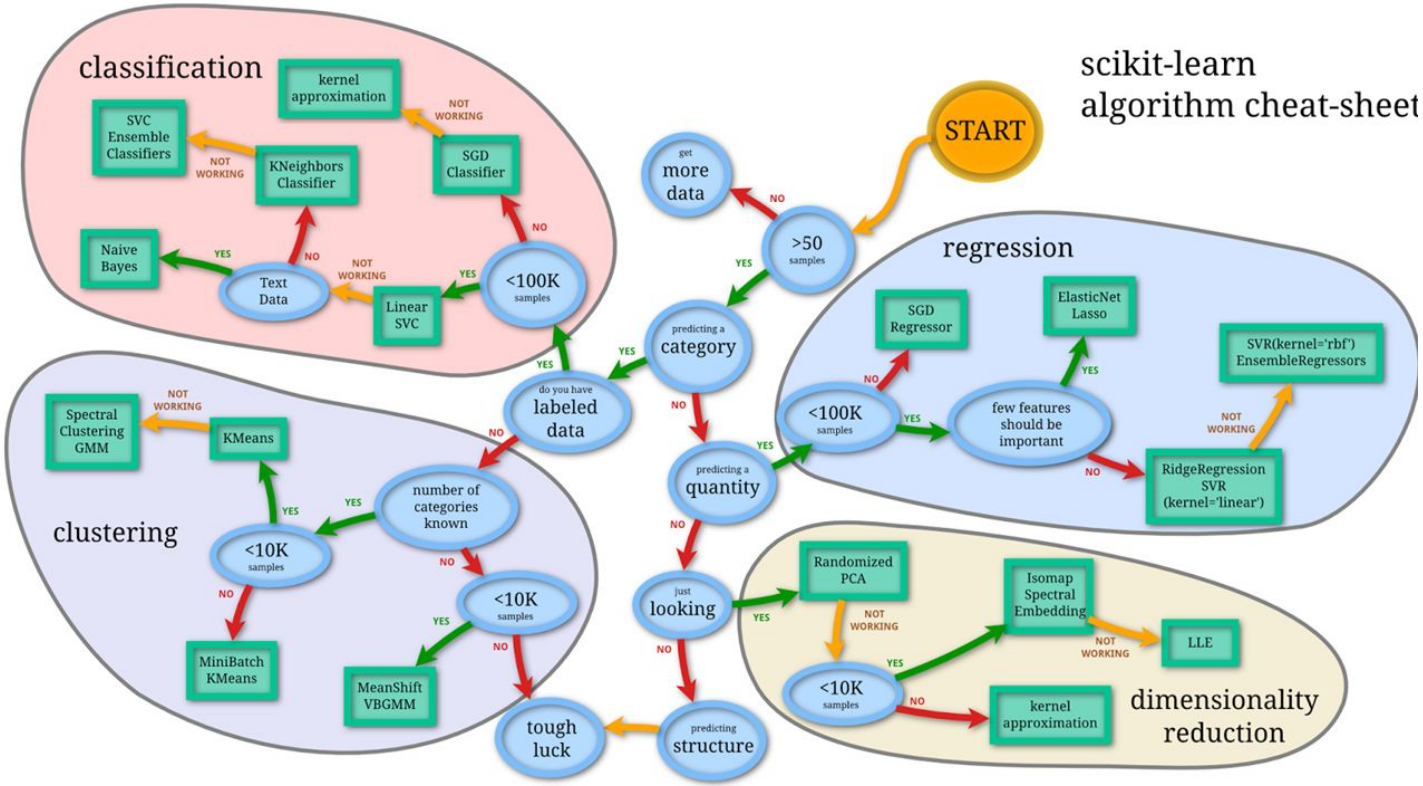
Storytelling

Storytelling

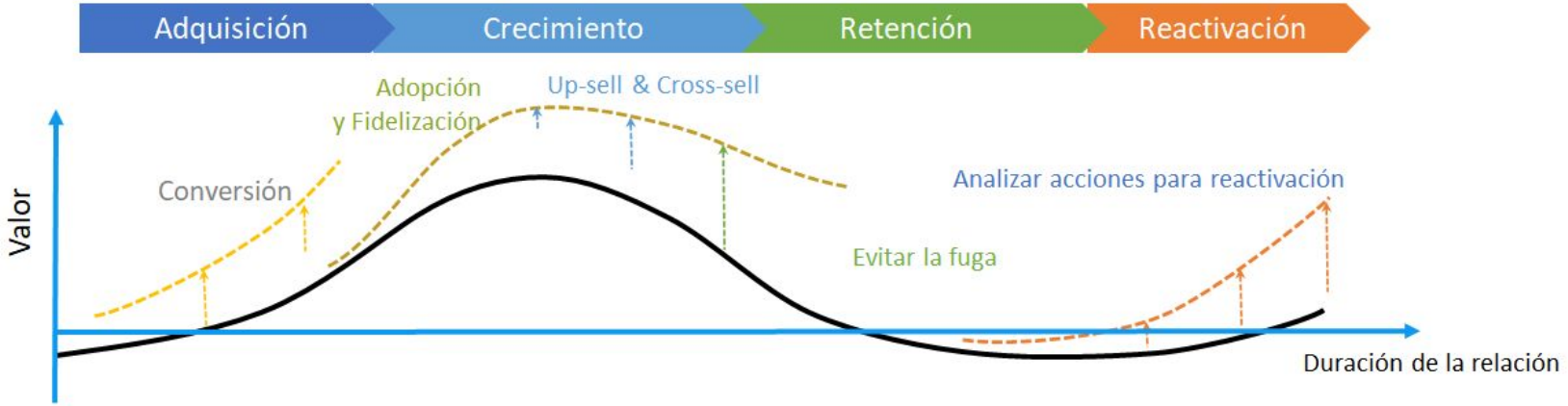
Juan Grande
Gerente de BI

Miranda Wintour
Directora Comercial

Hablando con el Negocio > Cómo piensa un Data Scientist



Hablando con el Negocio > Cómo piensa el negocio



Adquisición

- Look alike Models
- Audience Profile
- Real Time Offers

Crecimiento

- Segmentation (RFM)
- Cross-sell Modeling
- Up-sell Modeling
- Next Best Offer
- Recommendations
- Life time value

Retención

- **Attrition/Churn** Modeling
- Retention Offers
- Customer Value Analysis

Reactivation

- Look alike models
- Win-back offers

Primera Asignación > Abandono de clientes

Minuta de la reunión

- Nuestra empresa tiene clientes de alto valor que son los que disponen del Paquete Premium
- Un cliente de alto valor en promedio genera a la empresa 160k pesos al año.
- Adquirir a un cliente de alto valor es muy costoso, evitar que se vaya es más barato que conseguir uno nuevo.
- Se realizó un experimento, donde sí se gastaba 2000 pesos en un estímulo para retener a un cliente premium, el 50% acepta y se queda

Primera Asignación > Abandono de clientes

Minuta de la reunión

- Marketing quiere empezar a hacer campañas **proactivas** para evitar la fuga, **le pide a usted un listado de clientes que se estarían yendo a los cuales ellos deberían estimular.**
- Quieren la cantidad de clientes **“justa”**, les interesa **maximizar la ganancia**
- **NO** van a realizar acción alguna sobre el listado que le vayamos a dar. Van a comparar lo que enviamos con lo que paso en el real-world.

Primera Asignación > Ayuda de nuestro Jefe

Dado que Juan nos ve potencial decide ayudarnos explicandonos cómo el tiempo hace de un modelo de clasificación, un modelo predictivo. Foto de clientes y target futuro.

Foto de clientes

Toda info de los clientes para un mes particular. Solo tiene datos del pasado con respecto al mes

(4 dígitos para el año + 2 dígitos para el mes

ie: 202101)

Target

Si existe un cliente en una foto N meses adelantado del mes de la Foto

Valor:

CONTINUA

BAJA+1

BAJA+2

BAJA+N

Primera Asignación > Ayuda de nuestro Jefe

202010				202012
id	caja_ahorro	cheques	inversiones	clase
1	*	*	*	BAJA+2
2	*	*	*	CONT
3	*	*	*	BAJA+1
4	*	*	*	CONT
5	*	*	*	BAJA+1
6	*	*	*	CONT
7	*	*	*	CONT
8	*	*	*	BAJA+2

202011			
id	caja_ahorro	cheques	inversiones
1	*	*	*
2	*	*	*
4	*	*	*
6	*	*	*
7	*	*	*
8	*	*	*
9	*	*	*

202012			
id	caja_ahorro	cheques	inversiones
2	*	*	*
4	*	*	*
6	*	*	*
7	*	*	*
9	*	*	*
10	*	*	*

Primera Asignación > Ayuda de nuestro Jefe

		TARGET				
		202011	202012	202101	202102	202103
FOTO	202011		BAJA+1 CONTINUA	BAJA+2 BAJA+1 CONTINUA	BAJA+3 BAJA+2 BAJA+1 CONTINUA	BAJA+4 BAJA+3 BAJA+2 BAJA+1 CONTINUA
	202012			BAJA+1 CONTINUA	BAJA+2 BAJA+1 CONTINUA	BAJA+3 BAJA+2 BAJA+1 CONTINUA
	202101				BAJA+1 CONTINUA	BAJA+2 BAJA+1 CONTINUA

Primera Asignación > Ayuda de nuestro Jefe

Nos explica por qué hay que dejar un gap de un mes, y que solo tiene sentido predecir a dos meses, dejando la construcción final del target

		TARGET				
		202011	202012	202101	202102	202103
FOTO	202011		~~BAJA+1~~ ~~CONTINUA~~	BAJA+2 BAJA+1 CONTINUA	~~BAJA+3~~ ~~BAJA+2~~ ~~BAJA+1~~ ~~CONTINUA~~	~~BAJA+4~~ ~~BAJA+3~~ ~~BAJA+2~~ ~~BAJA+1~~ ~~CONTINUA~~
	202012			~~BAJA+1~~ ~~CONTINUA~~	BAJA+2 BAJA+1 CONTINUA	~~BAJA+3~~ ~~BAJA+2~~ ~~BAJA+1~~ ~~CONTINUA~~
	202101				~~BAJA+1~~ ~~CONTINUA~~	BAJA+2 BAJA+1 CONTINUA

Momento de reflexión - 5’

Este tema dado en sólo 4 slides es de suma importancia en su carrera. Grandes errores de diseño se han cometido por no saber construir adecuadamente un target.

- ¿Lo entendió claramente?
- ¿Qué información del futuro se puede añadir a la foto además del target?
- Sugiera la construcción de otro tipo de target.
- Intercambie sus interpretaciones con los compañeros de su mesa

Primera Asignación > Ayuda de nuestro Jefe

Nos explica cómo calcular la ganancia de una virtual campaña. Razonamiento:

- Debemos entregar una lista de personas que el modelo cree que no van a irse en 2 meses.
- A todas las personas en la lista habría que estimularlas con un coste de $2000, a priori no sabemos si se van, si se van en al mes siguiente o dentro de 2.
- Si la persona que dijimos que se iba en 2 meses, se va en el segundo mes, ganaríamos los $160000 del valor del cliente. Sin embargo como la experiencia dice que solo la mitad de los estimulados se quedan, decimos que si acertamos solo ganaremos 80000 pesos

Primera Asignación > Ayuda de nuestro Jefe

- Luego

  Ganancia = 0.5 * $160000 * acierto - $2000 * estimulos
  Ganancia= $80000 *$BAJA_2$ - $2000 *($BAJA_2$+ $BAJA_1$+ CONTINUA)

  Ganancia= $78000 *$BAJA_2$ - $2000 *($BAJA_1$+ CONTINUA)

- Ejemplo, enviamos 3 estímulos donde hay 1 $BAJA_2$, 1 $BAJA_1$y 1 CONTINUA
  - El $BAJA_2$ gasta $2000 en estímulo y ganaremos $80000. Tendremos una ganancia de $78000.
  - El $BAJA_1$ se nos habrá ido antes de tiempo, habremos perdido $2000.
  - El CONTINUA no se fue del banco. Habremos estimulado erróneamente por otros $2000.
  - La ganancia de esa campaña virtual de 3 casos nos dará $74.000

Primera Asignación > Ayuda de nuestro Jefe

Podemos representar la ecuación anterior a través de la siguiente matriz de ganancia

		Modelo		
		BAJA+2	BAJA+1	CONTINUA
Predicho	BAJA+2	78000	–2000	-2000
	BAJA+1	0	0	0
	CONTINUA	0	0	0

Momento de reflexión

		Real		
		BAJA+2	BAJA+1	CONTINUA
Predicho	BAJA+2	78000	–2000	-2000
	BAJA+1	0	0	0
	CONTINUA	0	0	0

- ¿Está bien definida la métrica de ganancia?
- ¿No debería haber un castigo cuando el modelo predice mal y decide no enviar estímulo a un registro que en realidad es un BAJA+2?

Primera Asignación > Cantidad de estímulos

Por último, ya teniendo cuál es la función de ganancias, tendremos que determinar la cantidad óptima de envíos para obtener la mayor ganancia.

Quizás algún alumnos piense en este momento:

> “¿Qué dice este? Los modelos nos devuelven la clase que corresponde, mandamos todos los que el modelo nos dice y listo.”

Bueno, NO. Un modelo devuelve un score, un valor habitualmente entre 1 y 0, que en algunos casos es la probabilidad de pertenencia de una determinada clase.

Haga un juego mental: 5’

- Si un cliente tiene una prob de 0.2 de ser BAJA+2, de 0.3 de ser BAJA+1 y de 0.5 de ser CONTINUA. ¿Usted le envía un estímulo?

Defienda su posición primero con los compañeros y luego con la clase.

Primera Asignación > Punto de corte

Dado que contamos con una función de ganancia, podemos definir cuales son todos los clientes que deben ser estimulados usando la probabilidad que devuelve el modelo.

$$f\left(x_i\right) = P\left(BAJA_2 | x_i\right) = p$$

f es el modelo que devuelve una probabilidad para el caso *x*

$$G\big|_{x_i} = \begin{cases} 78000 & p \\ -2000 & 1-p \end{cases}$$

G es una variable aleatoria (VA) que refleja la ganancia

$$E(G) = \sum_{g \in G} gP(G=g) = 78000p + -2000(1-p) = 80000p - 2000$$

$$E(G) \geq 0$$

$$-2000 + 80000p \geq 0$$

$$p \geq 0.025$$

Como es una VA, tiene una esperanza matemática (E) y si es positiva, refleja que ganamos plata con ese caso. Tan solo resta buscar el valor de *p* para el que la E es mayor a 0.

Luego todo caso cuyo modelo devuelva una p mayor a 0.025 será considerado bueno.