Welcome

Data Mining Economía y Finanzas - 2021 - Comisión 1

Alejandro Bolaños

Sobre Mí, Sobre Ustedes

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

Sobre esta Materia > Espíritu

Enseñar, más que llenar un recipiente es encender un fuego

El estilo en esta asignatura es el del aprendizaje basado en pares, en donde a pesar de existir un sendero diseñado por el profesor se alienta continuamente a los alumnos a recorrer su propio camino, observar y aprender de sus pares, diseñar creativamente sus propios experimentos reflexionando profundamente sobre los resultados que se van obteniendo, cuestionando al establishment.

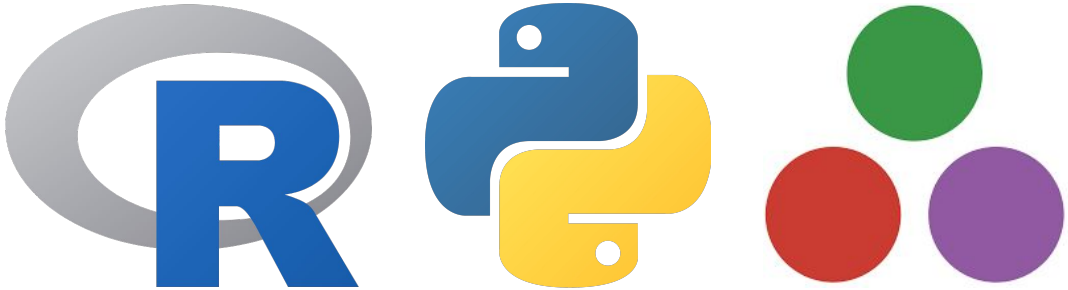
Sobre esta Materia > OBJETIVOS DE LA MATERIA

- 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.
- Conocer y utilizar efectivamente las técnicas “estado del arte” de algoritmos y librerías de última generación de modelado predictivo sobre datos estructurados.

Sobre esta Materia > ¿Qué vamos a ver?

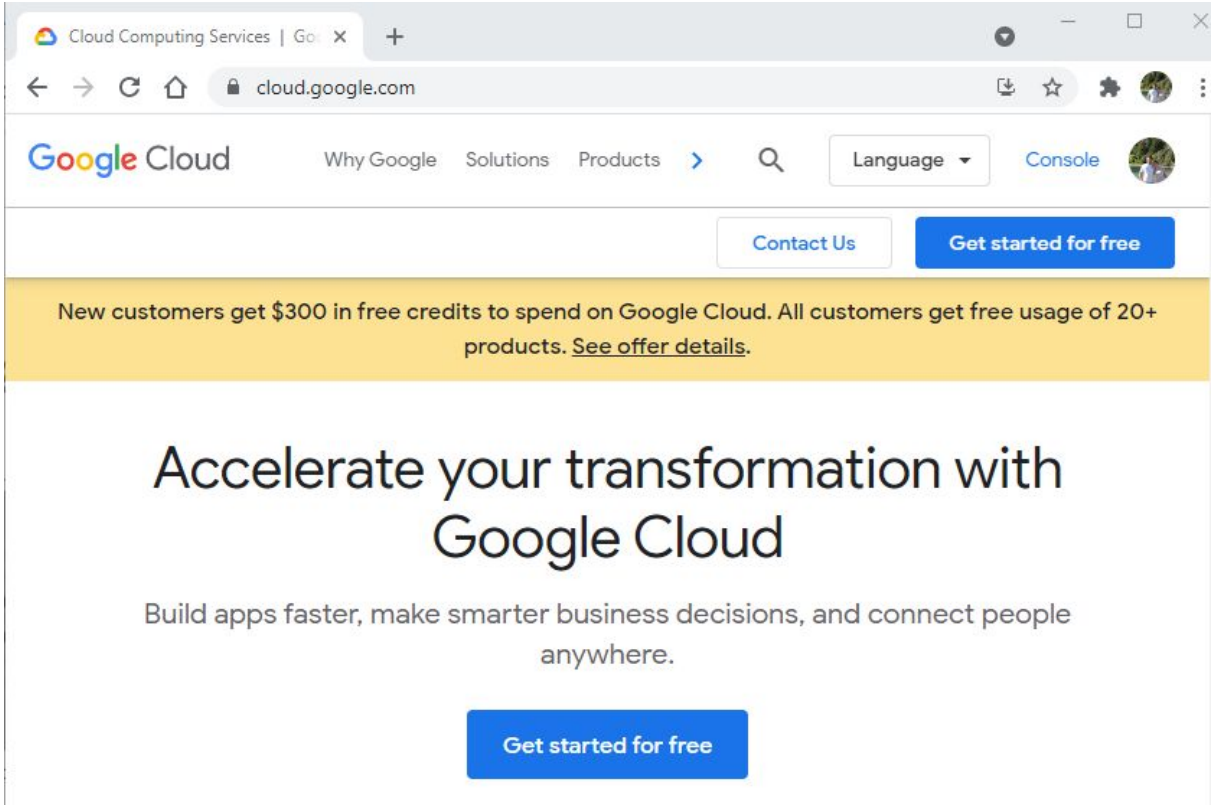
- Metodologías
- Algoritmos y librerías de última generación de modelado predictivo sobre datos estructurados.
- Feature Engineering
- Optimización de Hiperparámetros
- Explicabilidad de modelos black box.
- MLOps
- Big Data en la nube

Sobre esta Materia > Lenguajes de Programación



- Breve historia de los lenguajes
- Adopción
- Para el mundo de la ciencia de datos
- Pros y Contras

Sobre esta Materia > Nube



En un momento de la cursada, sus computadoras personales no les alcanzarán para procesar los modelos

Sobre esta Materia > Fechas

Comisión		Hora	Actividad
Lunes	Jueves		
18-ago		04:30	Envio por email de este documento a alumnos
18-ago	19-ago	19:00 a 22:00	Primera Clase
13-sep	09-sep	21:00	Lanzamiento Actividad sobre Feature Engineering "Dos Universidades"
20-sep	23-sep	19:00	Habilitación Scripts instalación Google Cloud
mie 29-sep		19:00	Lanzamiento "Desafío Cazatalentos"
dom 17-oct		23:59	Cierre automático de la Primer Competencia Kaggle
lun 18-oct		18:00	Lanzamiento Segunda Comptencia Kaggle
18-oct	21-oct	19:00	Análisis de resultados de la Primer Competencia, ¿Cómo evitar overfittear el Public Leaderboard? Lanzamiento lenguaje Julia
05-nov		19:00 a 22:00	Clase Conjunta Repensando el overfitting en 2021 con nuevo código para GBDT

mie 01-dic		23:58	Fecha límite entrega de los dos Videos y el brevísimo informe
		23:59	Cierre automático de la Segunda Competencia Kaggle
jue 02-dic		19:00 a 22:30	Presentación de equipos ganadores Segunda Competencia
jue 02-dic		20:30 a 22:00	Evaluación Individual Escrita con constancia para la CONEAU
vie 03-dic		23:00	Lanzamiento actividad Torneo de Videos
mie 08-dic		23:59	Fin actividad Torneo de Videos
06-dic	09-dic	19:00 a 22:00	Ultima Clase, Recapitulación, Lecciones Aprendidas, Mejores Prácticas, ¿Bala de Plata?
dom 12-dic		23:00	Entrega de notas por parte de los profesores. Se determina quienes pasan a recuperatorio.
mie 15-dic		23:00	Entrega de instrucciones, datasets y scripts para recuperatorios individuales
30-abr-2022		23:59	Fecha límite de entrega de recuperatorio

Sobre esta Materia > Hitos y Notas

Contribución Nota	Actividades Obligatorias
60%	Segunda Competencia Kaggle más brevísimo reporte. Una o dos personas.
15%	Primera Competencia Kaggle - Una o dos personas.
15%	Participación significativa en Zulip + hypothes.is
5%	Video Presentación de 5 minutos dirigido al Gerente de Business Intelligence - Individual
5%	Video Presentación de 5 minutos con storytelling a la Directora Comercial - Individual

Nota adicional	Actividad
+1	Cazatalentos 15k
+2	Cazatalentos 14k (no acumulable)

Ver cronograma en el Libro de la Materia

Herramientas > Zulip

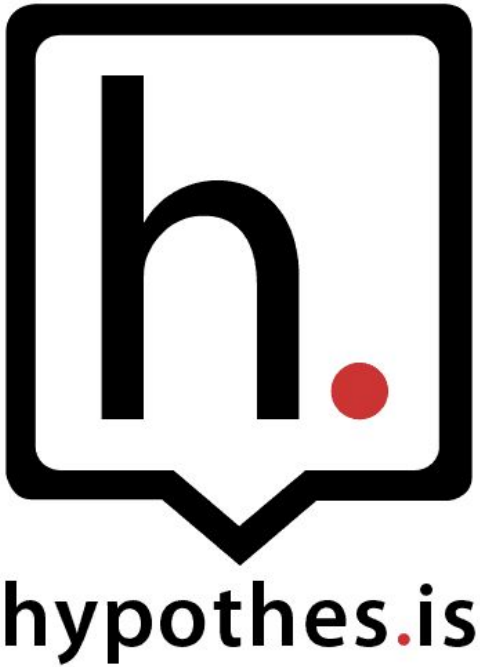
https://dmeyf2021.zulip.rebelare.com/

MUY IMPORTANTE: NO SE RESPONDEN MAILS DE CONSULTA

Herramientas > Zulip > Nota

Rúbrica de participación en foro Zulip	
Porcentaje	**Concepto a Evaluar**
30%	Contenido. Propone ideas profundas, signficativas, claras, fáciles de implementar y con potencial para el proyecto de la asignatura. Compartiendo resultados logra atraer la atención y colaboración de sus compañeros a sus posts. Idealmente se transforma en un líder conceptual del grupo.
30%	Contribución al dinamismo de la comunidad. Plantea preguntas interesantes y relevantes, intenta motivar las discusiones grupales sobre tópicos relevantes a la asignatura, debate positivamente, participa de conversaciones iniciadas por otros.
20%	Colaboración en dudas operativas, colabora rápida y acertadamente con compañeros que requieren algun tipo de asistencia operativa en alguna de las herramientas del curso, lenguaje de programación, etc Idealmente se transforma en un referente en un tema específico del grupo.
20%	Frecuencia. Todas las semanas posee relevantes participaciones.

Herramientas > hypothes.is



https://hypothes.is/groups/e3anZ53M/labo2021
https://hypothes.is/groups/KoEJYniw/labo2021-art
https://hypothes.is/groups/edgz1vNd/labo2021-code

Herramientas > hypothes.is > Primer Tarea

- https://storage.googleapis.com/dmeyf/annotation/general/why_so_many_data_scientists_are_leaving_the ir_jobs.html
- https://storage.googleapis.com/dmeyf/annotation/general/why_business_fail_at_machine_learning.html
- https://storage.googleapis.com/dmeyf/annotation/code/101_PrimerModelo.R.html
- https://storage.googleapis.com/dmeyf/annotation/code/101_PrimerModelo.html

Herramientas > hypothes.is > Nota

Rúbrica de participación en Anotacion Colaborativa	
Porcentaje	**Concepto a Evaluar**
30%	Profundidad y originalidad de la Interpretación, la mayoría de los comentarios del alumno revelan que ha evaluado el documento en detalle y reflexionado profundamente sobre el significado, aportando ideas originales
30%	Participación en conversaciones (responder anotaciones de otros alumnos), el alumno agrega sustancia a los comentarios de compañeros, más alla de simplemente indicar aprobación o desaprobación
20%	Cantidad de comentarios, el alumno agrega varios comentarios relevantes en cada uno de los documentos.
20%	Clarificación de conceptos, el alumno colabora activamente en clarificar conceptos relevantes, propone links interesantes, relaciona conceptos complejos.

Herramientas > Videos (no mandatorias)

- Windows Movie Maker
- iMovie
- OpenShot
- Adobe Spark
- Da Vinci
- Vimeo
- Canva
- Youtube
- TikTok
- .....

Herramientas > Videos > Notas

Rúbrica Videos Presentación	
Porcentaje	**Concepto a Evaluar**
10%	Entretenimiento. El video presentación es apasionante y no hay parte que aburra, son 5 minutos de pura adrenalina.
10%	Audiencia. El video presentación tiene totalmente en cuenta la audiencia para la cual está dirigido y saca provecho de las características únicas de esa audiencia.
30%	Historia . La presentación narra una historia, hay una clara introducción con un "gancho" que invita a ver el video, un desarrollo adecuado con una continuidad argumental lógica y un desenlace concreto. La narrativa está organizada en torno a las etapas de la pirámide de Freytag o estructura de similar complejidad donde la emoción juega un papel fundamental.
25%	Consistencia del contenido. Lo presentado refleja fielmente el conocimiento descubierto en la Segunda Competencia Kaggle y las conclusiones están sustentadas en datos que aparecen presentados adecuadamente.
25%	Originalidad del contenido. Las ideas presentadas son originales, ingeniosas, basadas en una profunda comprensión del problema.

Herramientas > Kaggle + Github

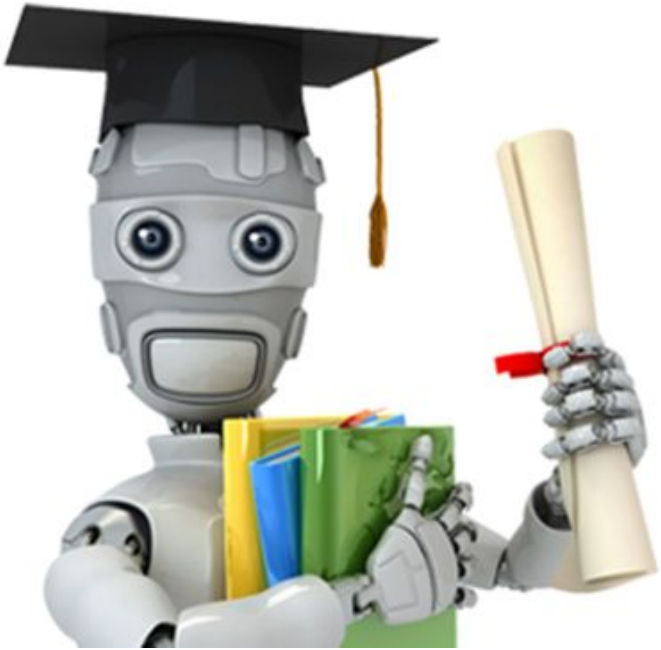
https://www.kaggle.com/c/uba-dmeyf2021-primera/

Link Invitación a la Competencia Kaggle
https://www.kaggle.com/t/2410cfcc7d804b71a5b176231a442d39

https://github.com/dmecoyfin/dmeyf

Break

Storytelling



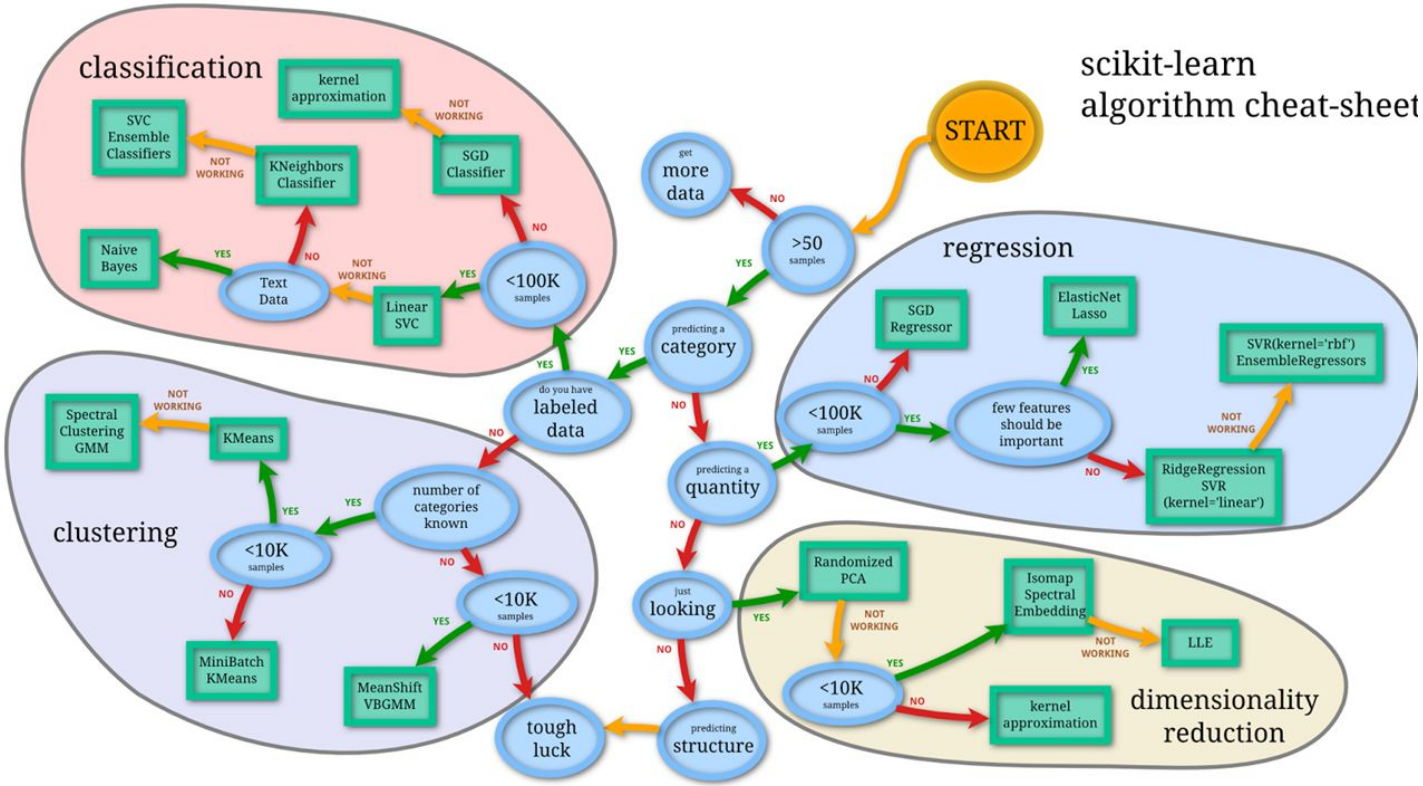
Storytelling

Storytelling

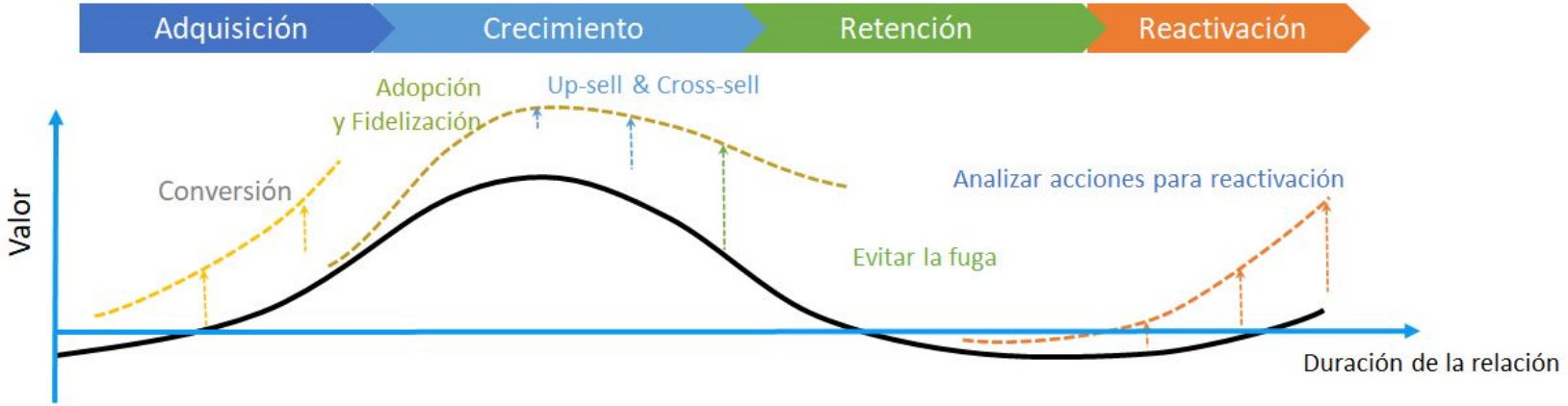
Juan Grande
Gerente de Ciencia de Datos

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

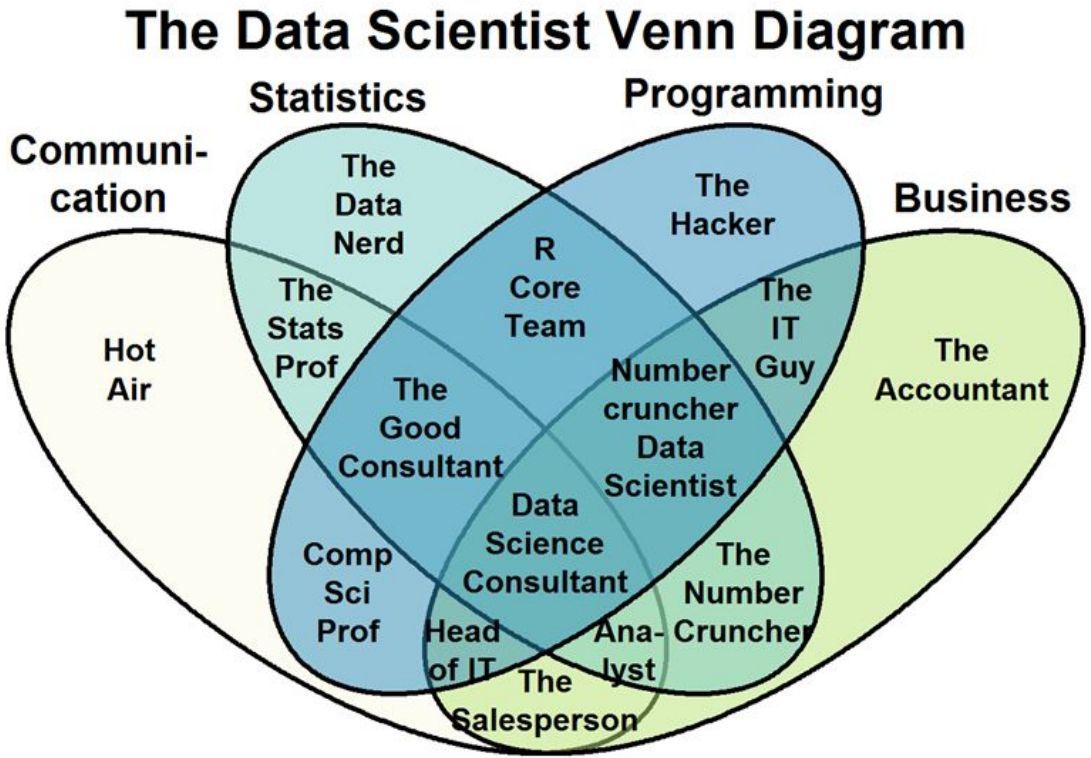
Retención

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

Reactivation

- Look alike models
- Win-back offers

Storytelling > Usted, un unicornio



Primera Asignación > Abandono de clientes

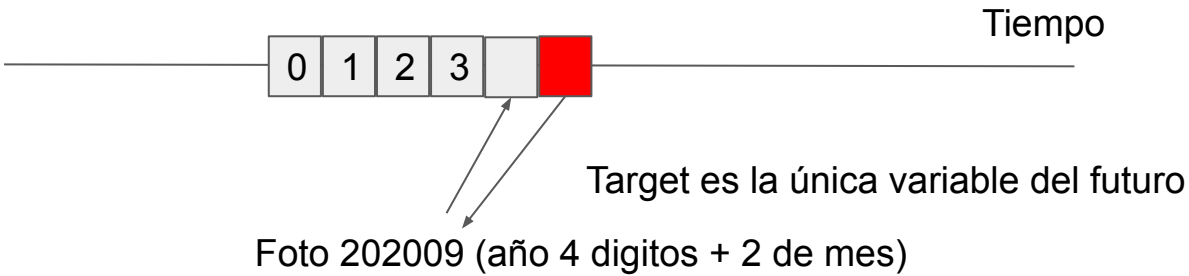
Minuta de la reunión

- 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 100k pesos
- Adquirir a un cliente de alto valor es muy costoso
- Se realizó un experimento, donde sí se gastaba 1250 pesos en un estímulo para retener a un cliente premium, el 50% acepta y se queda
- Marketing quiere empezar a hacer campañas proactivas para evitar la fuga, le pide un listado de clientes a los cuales ellos deben estimular.
- Quieren la cantidad de clientes justa, les interesa maximizar la ganancia

Primera Asignación > Ayuda de nuestro Jefe

Dado que Jorge nos ve potencial decide ayudarnos de la siguiente manera:

Nos explica cómo el tiempo hace de un modelo de clasificación, un modelo predictivo. Foto de clientes y target futuro.



Target: Se fue | No se fue

Primera Asignación > Ayuda de nuestro Jefe

Nos explica por qué hay que dejar un gap y nos muestra la construcción de las clases

Primera Asignación > Ayuda de nuestro Jefe

Nos explica cómo construir la función de ganancia.

		Predicho	
		BAJA+2	BAJA+1, CONTINUA
Real	BAJA+2	48750	0
	BAJA+1, CONTINUA	-1250	0

Primera Asignación > Ayuda de nuestro Jefe

Nos da los datos! y nos hace una importante advertencia: La empresa no pierde muchos clientes premium por mes, tan solo 0,35%.

¿Cómo puede afectar esto a nuestro modelo?

¿Cómo determina la clase una librería de Machine Learning?

¿Cuán frecuente cree que es este escenario?

Primera Asignación > Punto de corte

Y como último favor, no ayuda a calcular cuál es el punto de corte para determinar la clase.

$$f(x_i) = P(B + 2|_{x_i})) = p$$

$$G|_{x_i} = \begin{cases} 48750 & p \\ -1250 & 1 - p \end{cases}$$

$$\begin{aligned} E(G) &= \sum_{g \in G} gP(G = g) \\ &= 48750p - 1250(1 - p) \\ &= -1250 + 50000p \end{aligned}$$

$$\begin{aligned} E(G) &\geqslant 0 \\ -1250 + 50000p &\geqslant 0 \\ p &\geqslant 0.025 \end{aligned}$$

Retorno a la realidad

> Veamos los scripts en Hypothes.is
> Navegemos por Kaggle

> Tarea