OCT 2014

Data Mining y los sesgos de selección

Discusión de casos: Originación de Créditos y Detección del Hurto de Electricidad

IX Jornadas de Data Mining y Business Intelligence
Tenaris University, 17/10/2014

Fernando Castelpoggi

BeSmart.
SMART BUSINESS SOLUTIONS

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

AGENDA: Data Mining y los sesgos de selección

BeSmart. SMART BUSINESS SOLUTIONS

- Objetivos de la charla
- Debate: Minería de Datos contempla tratar sesgos de selección?
- Caso 1: Modelos para la Originación de Créditos
- Caso 2: Modelos para la Detección de Hurto de electricidad
- Discusión y recomendaciones

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Smart Business Solutions S.A.

BeSmart. SMART BUSINESS SOLUTIONS

- Empresa dedicada a la provisión de Soluciones inteligentes (Software, Servicios, Capacitación y Consultoría).
- Desde hace más de 15 años en Argentina y Chile.
- Partners de IBM: SPSS Modeler, Unica, Ilog, Tealeaf, Open Pages, etc.
- Profesionales con amplia experiencia aplicada a diversas industrias y procesos.
- Certificación ISO-9001 de los procesos de desarrollo e implementación.

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Objetivos

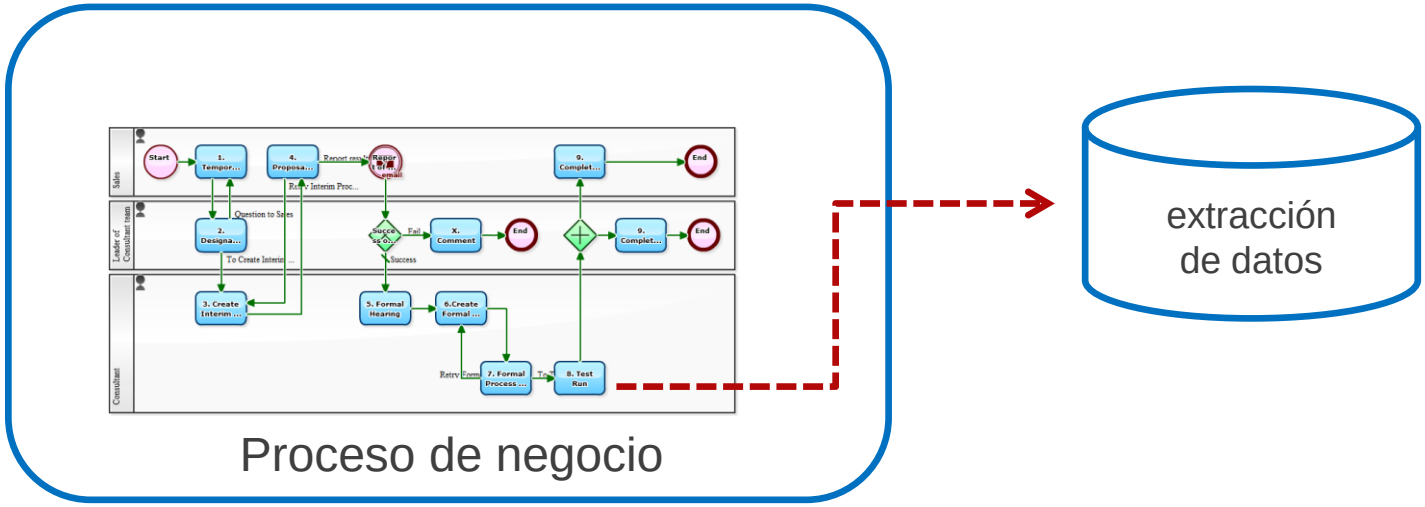
BeSmart.
SMART BUSINESS SOLUTIONS

- Evaluar efectos de la presencia de sesgos de selección en el contexto actual de uso intensivo de modelos y gran disponibilidad de datos (Big Data).
- Desarrollar intuición para discernir si el proceso que generó nuestro set de datos presenta sesgo de selección no despreciable.
- Para ello, analizaremos dos ejemplos de aplicación:
  - **Modelos de Scoring para Originación de Créditos**
  - **Modelos para la Detección de hurto de electricidad (fraude)**
- Implicancias prácticas en caso de que nuestra muestra de datos presente un sesgo de selección importante:
  - Diseñar un plan de proyecto adecuado
  - Evaluar si el impacto es el mismo en todos los segmentos de interés
  - Conocer en qué consisten algunos métodos de inferencia que se pueden aplicar para contrarrestar el sesgo
  - Implicancias para el seguimiento de los modelos
  - Planificar, en la medida de lo posible, un diseño experimental (grupos de testeo / control) del proceso que genera los datos, adecuado para futuros análisis / modelos.

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Contexto

- Sesgo de selección: cuando la muestra disponible no es representativa de la población de interés (*).
- En el contexto de Minería de Datos, esta situación se suele dar en dos situaciones:
  - cuando el proceso que generó los datos no fue parte de un diseño experimental, ó
  - cuando el costo de ampliar la muestra es muy elevado



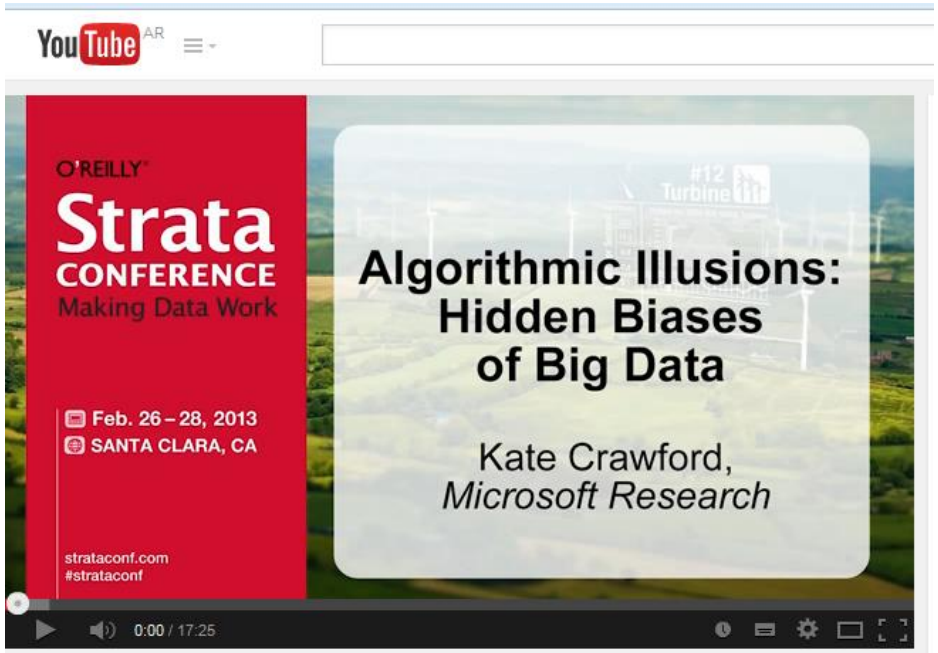
Los procesos de negocio tienden a sesgan los datos que luego serán analizados, excepto cuando hubo un diseño con fines analíticos. Pero esto requiere una fuerte cultura analítica, ya que implica implementar estrategias con grupos de testeo y control (además de hacer un seguimiento adecuado)

(*) definición no técnica

Contexto: Big Data (y Big Biases)

BeSmart.
SMART BUSINESS SOLUTIONS

- Abundancia de datos no garantiza la ausencia de sesgos, si los procesos que los generan no fueron parte de un diseño experimental
- En este contexto, la excepción es encontrar muestras insesgadas, y se trata de saber discernir situaciones que requieren especial atención



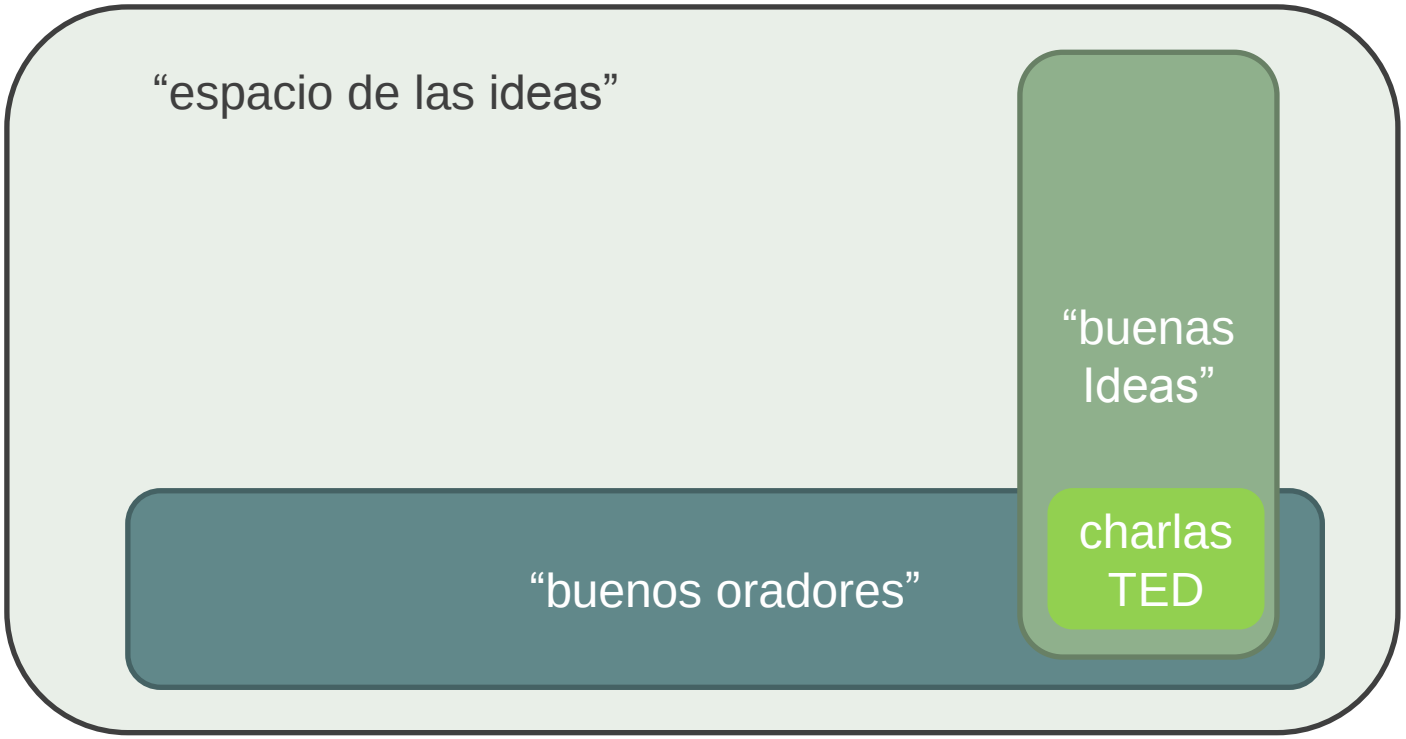
Strata 2013: Kate Crawford, "Algorithmic Illusions: Hidden Biases of Big Data"

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Ejemplo: sesgo en las charlas TED?

BeSmart.
SMART BUSINESS SOLUTIONS

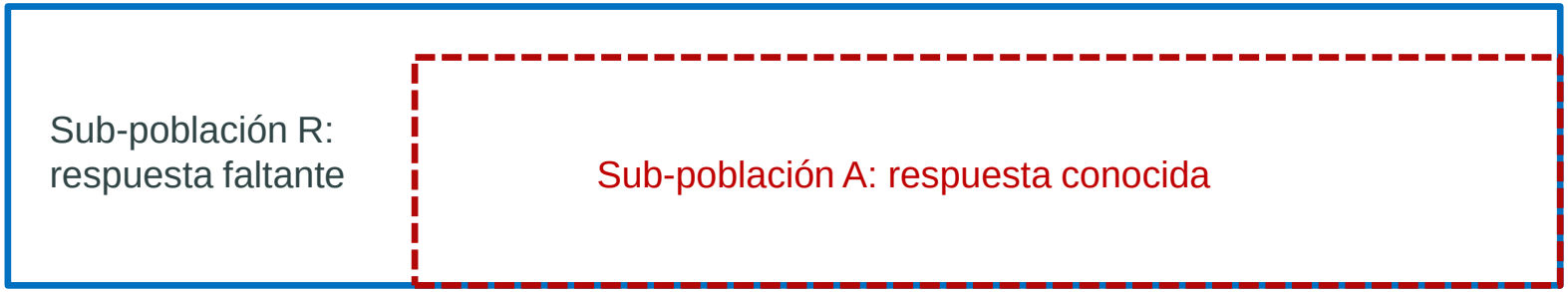
Las charlas TED se proponen muestrear el subconjunto de las “buenas ideas”, pero dicha muestra podría estár sesgada…ya que se debe restringir a los buenos oradores.



UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

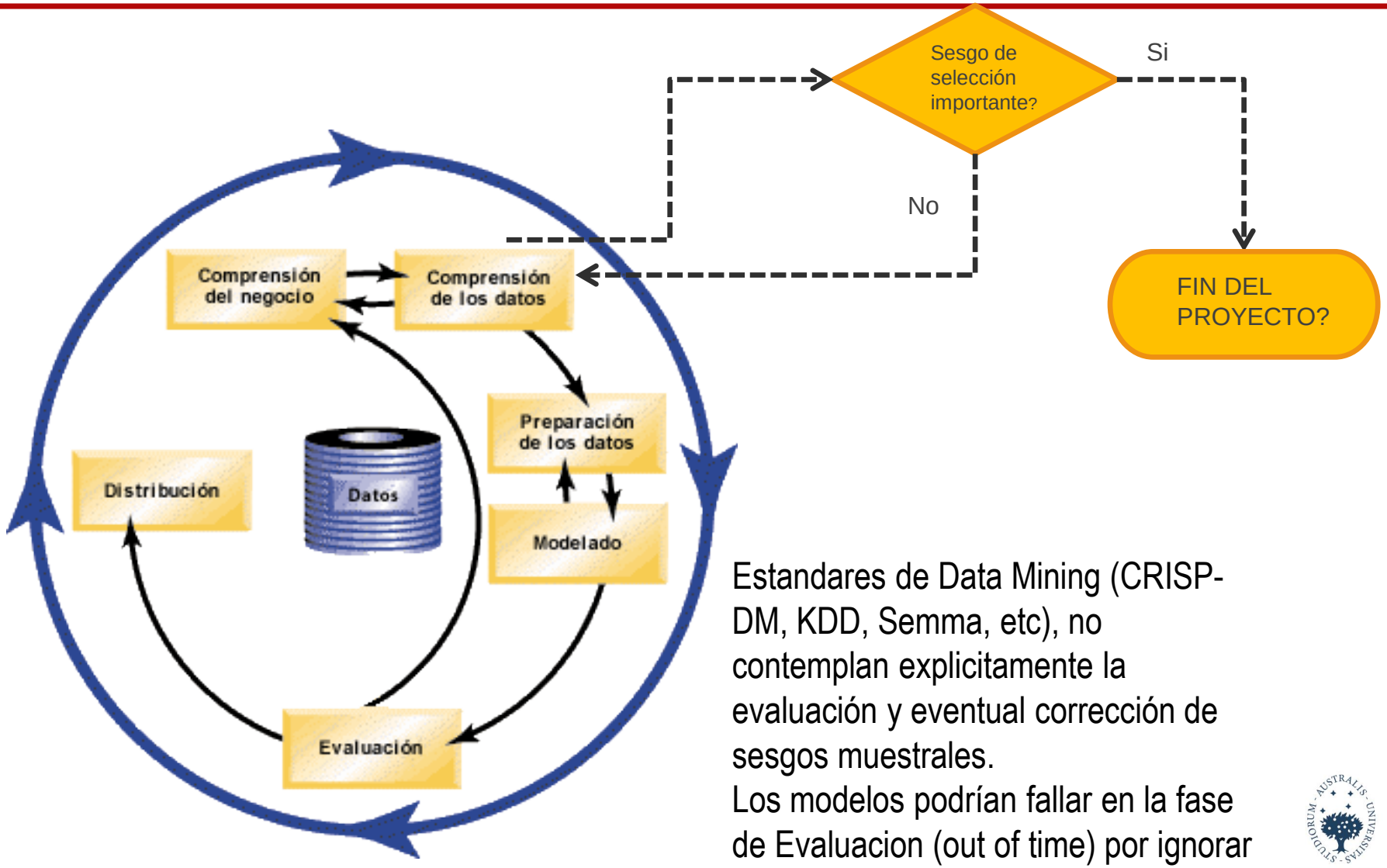
Sesgos

- Hay muchas variantes de sesgo de selección
  - Sesgo de intervalo de tiempo: un estudio temporal concluye antes de tiempo (cuando los resultados soportan una determinada conclusión). El sesgo de supervivencia es un caso particular.
  - De información: datos son descartados con criterio arbitrarios o no estadisticos
  - Sesgos de muestreo no aleatorio
- En particular, nos focalizaremos en la siguiente situación:
  - Queremos obtener modelos predictivos (aprendizaje supervisado) a partir de una muestra de datos, pero existe una subpoblación para la cual la variable target (respuesta) no se conoce.
  - La ausencia de respuesta (sub-población R) no es aleatoria, sino que está relacionada a lo que queremos predecir. Ejemplo: Créditos Aprobados (A) y Rechazados (R)



Data Mining y los sesgos de selección

BeSmart.
SMART BUSINESS SOLUTIONS

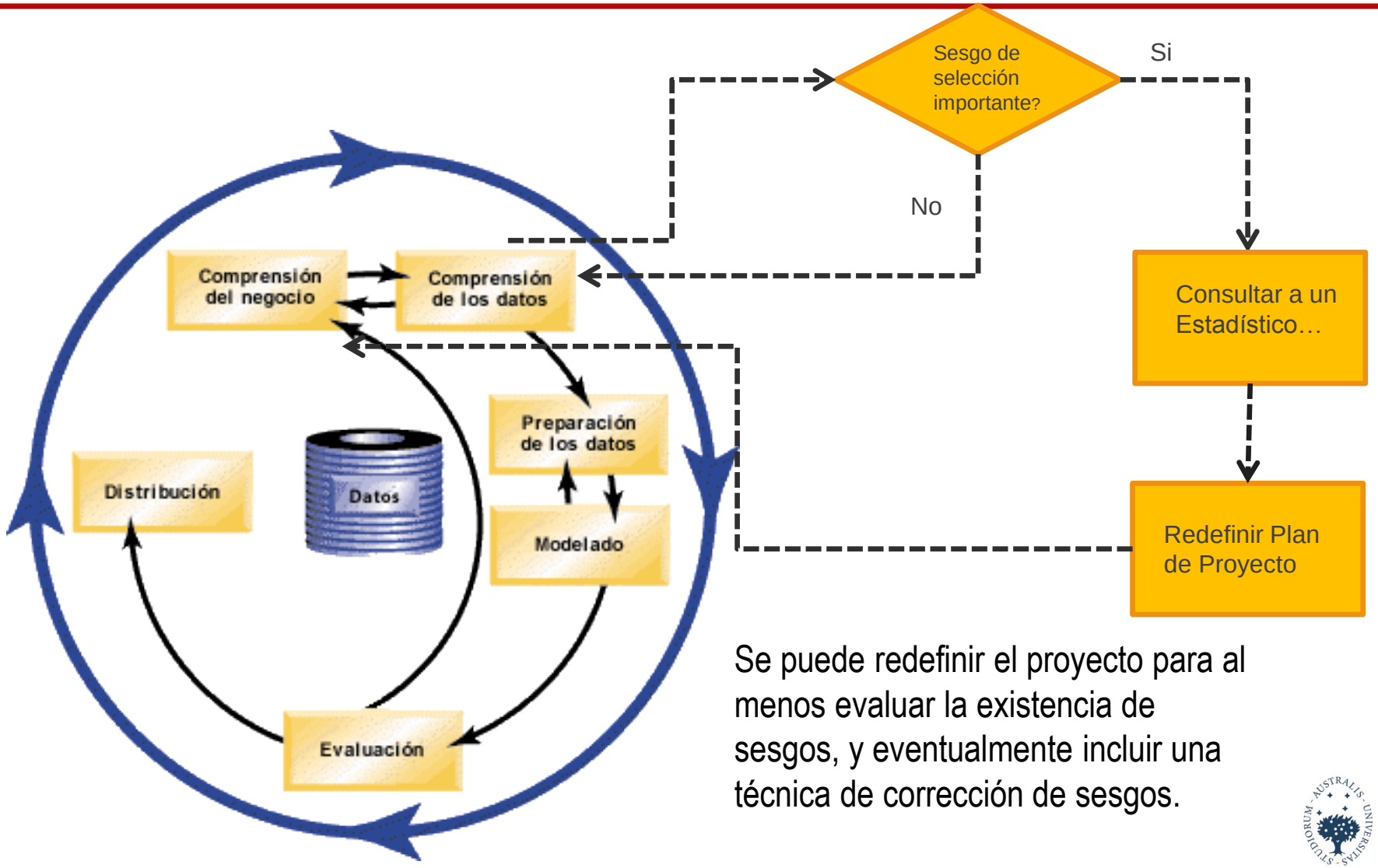


Estandares de Data Mining (CRISP-DM, KDD, Semma, etc), no contemplan explicitamente la evaluación y eventual corrección de sesgos muestrales.
Los modelos podrían fallar en la fase de Evaluacion (out of time) por ignorar sesgos importantes

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Data Mining y los sesgos de selección

BeSmart.
SMART BUSINESS SOLUTIONS



Se puede redefinir el proyecto para al menos evaluar la existencia de sesgos, y eventualmente incluir una técnica de corrección de sesgos.

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

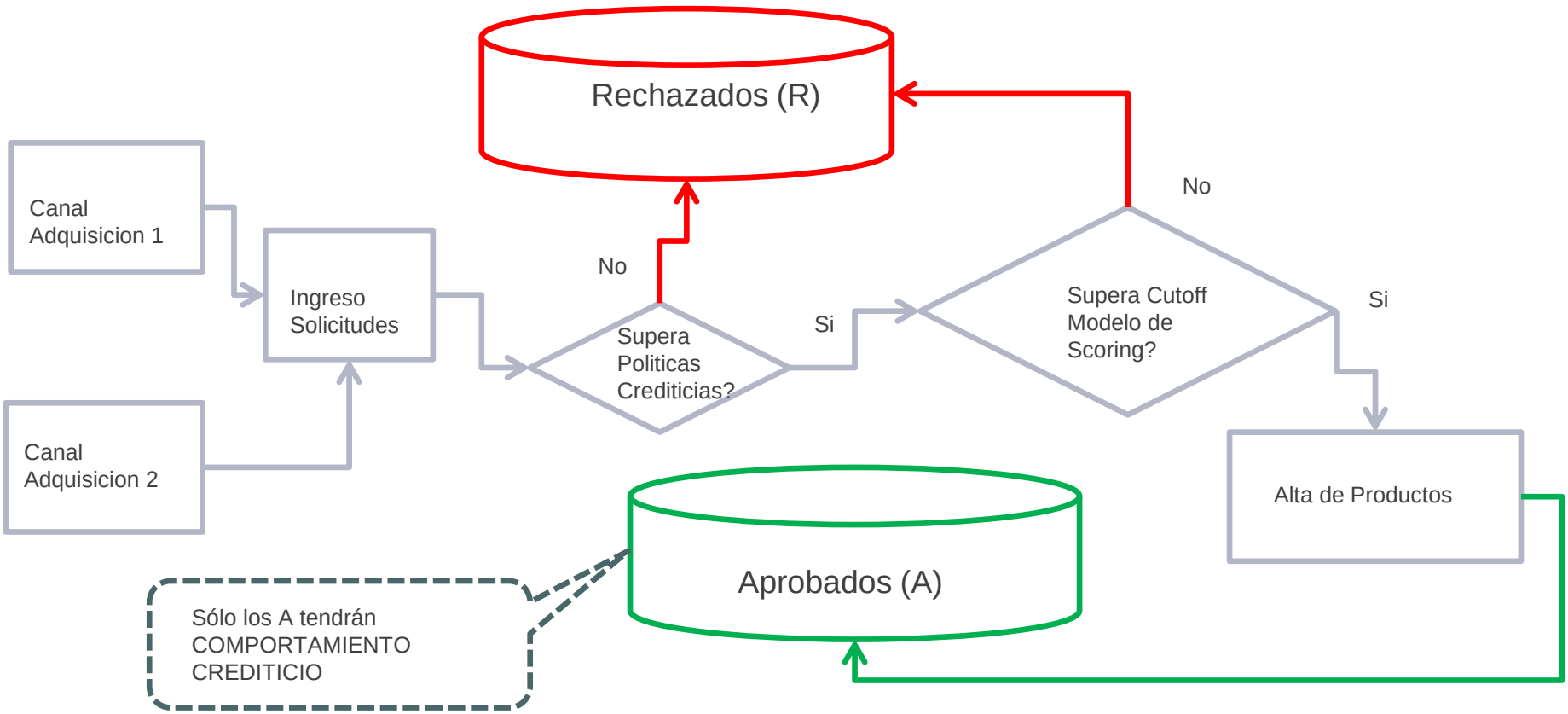
BeSmart.
SMART BUSINESS SOLUTIONS

Caso 1: Modelos de Scoring - Originación de Créditos

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Originación de Créditos

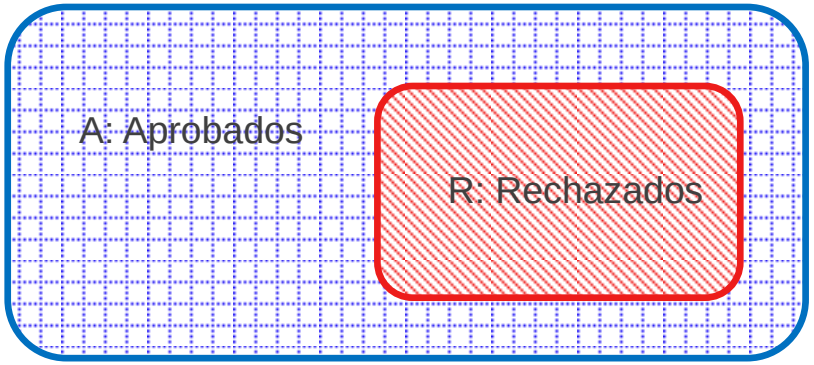
BeSmart.
SMART BUSINESS SOLUTIONS



En este workflow simplificado, vemos que los Aprobados generarán comportamiento crediticio (bueno o malo, el target de nuestro modelo), mientras que los Rechazados no Al momento de generar un nuevo modelo de Scoring, el sesgo se producirá al utilizar una muestra de Aprobados.

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

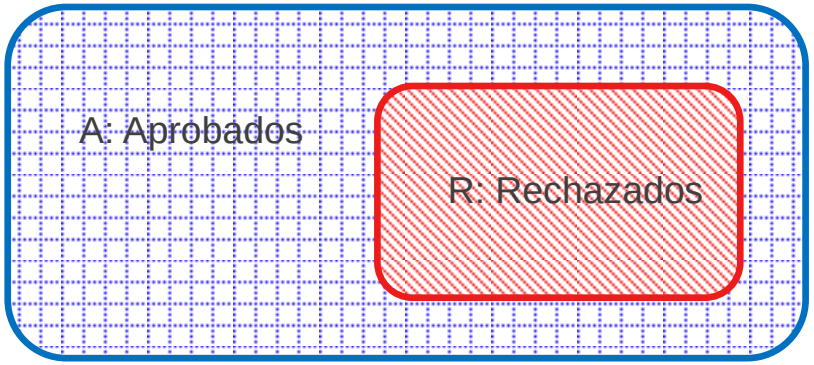
Medida del Sesgo



- Qué pasa si sólo trabajamos con los Aprobados?
  - El efecto depende del tamaño del sesgo
- El tamaño del sesgo estará dado en general por:

1. La eficiencia de las reglas / modelos que produjeron los rechazos:
   - En el extremo en que es aleatorio, el sesgo es nulo.
   - En el otro extremo (predicción perfecta), bastará con asignar a todos los R como malos pagadores.
   - La situación intermedia será lo habitual
2. La tasa de rechazo:
   - A mayor tasa de rechazo, mayor será el impacto de una posible inferencia de los R
3. El ratio entre la tasa de rechazos y la tasa de eventos en A (Malos / Aprobados):
   - Cuanto mayor es este ratio, más impacto tendrá la inferencia de los R.

Efectos del Sesgo



- Qué pasa si sólo trabajamos con los Aprobados?

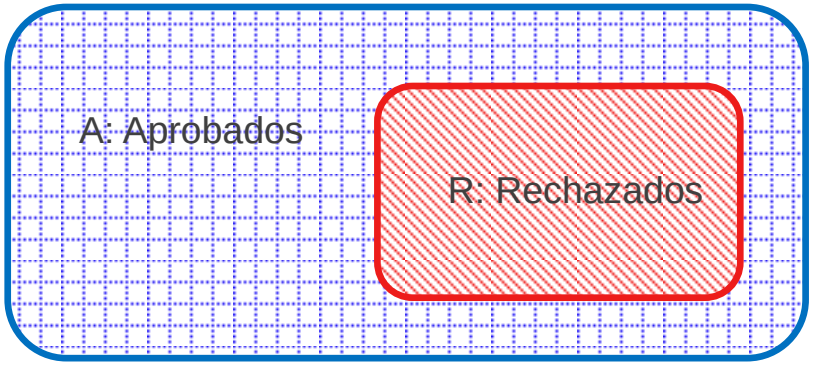
1. Habrá una segura subestimación en las estimaciones de tasa de default (total y por rangos de score)
2. Eventual efecto en la selección de variables por efectos distorsivos en los patrones de las variables.

   Se denomina efecto “cherry picking” (*) cuando dentro de los criterios de rechazo se aplicó una política muy selectiva a un grupo considerado de alto riesgo.

   Ejemplo: Supongamos que se sabe que la tasa de morosidad de los motoqueros es muy alta. Por ello se les requiere mucha documentación y avales y eso lleva a una tasa de rechazo del 90%. El 10% aprobado podría mostrar un comportamiento mucho mejor a la media de aprobados, y el modelo podría ponderar favorablemente la característica “Es_Motoquero” .

(*) En español: falacia de la prueba incompleta

Inferencia de Rechazos



- Los métodos de inferencia de rechazos apuntan a reconstruir una muestra que refleje el total poblacional.
- Hay un número importante de métodos de Inferencia de Rechazos y abundante literatura al respecto
- Razones para efectuar inferencia de Rechazos
  - Corregir el sesgo y los posibles efectos en la seleccion de variables
  - Reflejar la información de las políticas de rechazo aplicadas en el pasado
  - Permite construir un “swap-set analysis” en el que se simula una comparación entre los aprobados y rechazados por el modelo previo vs el nuevo modelo,
  - Permite encarar otros análisis de tipo “what if” (reducción de puntos de corte –cutoff-)
  - Mejor estimación de tasas de interés ajustadas al riesgo

Métodos de Inferencia de Rechazos

BeSmart.
SMART BUSINESS SOLUTIONS

- Describiremos algunos métodos de Inferencia de Rechazos a modo introductorio.
  - Nota previa: se supone que la variable respuesta o target para los A puede tomar los valores G (buenos pagadores) o B (malos pagadores) con un criterio bien definido sobre un período de performance dado.

  Métodos “triviales”:
  - List-wise deletion: ignorar los R (es decir, no hacer nada)
  - Asignar respuesta de todos los R como B
  - Asignar lor R al azar G y B en función de la tasa de evento sobre A (la tasa de evento es B/(B+G))

  Hasta aquí, estos métodos no requieren contar con variables independientes (predictores de la variable target) para los Rechazados; esto será necesario para los siguientes métodos.

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Métodos de Inferencia de Rechazos

BeSmart. SMART BUSINESS SOLUTIONS

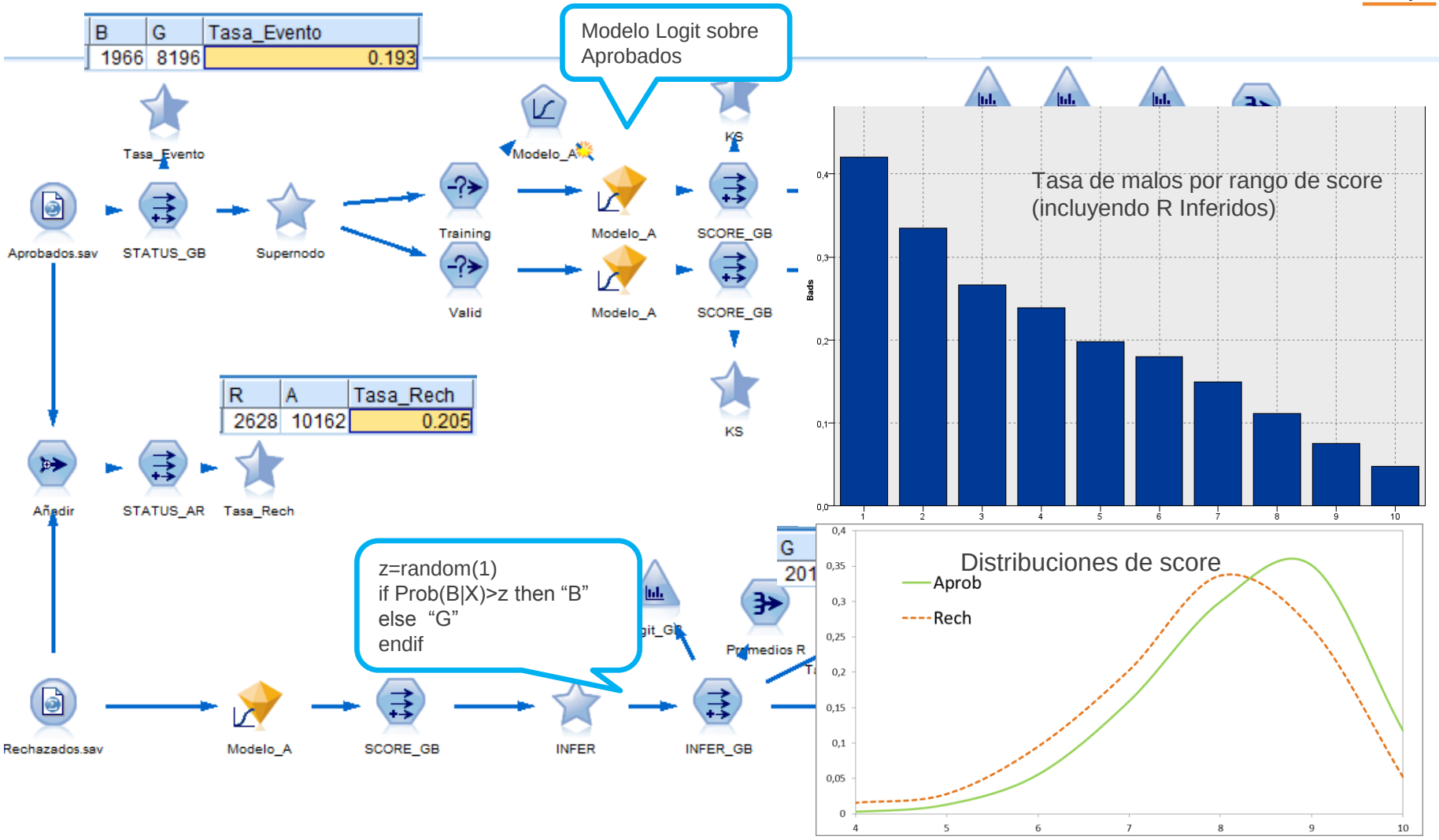
Método (mínimo) para corregir estimaciones poblacionales:

- Parceling:
  - Se obtiene un modelo M_A sobre Aprobados, a partir del cual se pueda obtener una estimación de la probabilidad de evento P(B|X), siendo X las variables utilizadas por el modelo.
  - Se aplica M_A a los Rechazados y se estima su P(B|X)
  - Con probabilidad P(B|X) se incorpora a cada R como B ó G.
  - Se obtiene la muestra aumentada A+R sobre la cual se pueden obtener estadísticas poblacionales corregidas, estadísticas por rango de score, swap-set analysis, etc. También se suele reestimar el modelo, pero su efecto, por las características de la imputación, es limitado.
- Otros métodos de mayor complejidad:
- Augmentation (varias variantes alrededor de esto)
  - Se modela el proceso de aprobación, es decir, se obtiene un modelo en el que el target es el estado A / R (este modelo).
  - Dicho modelo debe proveernos una estimacion de la P(A|X), siendo X variables predictivas
  - Se ponderan los Aprobados con un peso inverso a la probabilidad de aprobación P(A|X), y se construye un modelo sobre A con dichas ponderaciones

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Método de Parceling

Plantilla de IBM SPSS Modeler



Métodos de Inferencia de Rechazos

BeSmart. SMART BUSINESS SOLUTIONS

- Otros métodos de mayor complejidad (cont):
  - Fuzzy Augmentation
    - Se obtiene un modelo M_A sobre Aprobados, a partir del cual se pueda obtener una estimación de la probabilidad de evento P(B|X), siendo X las variables utilizadas por el modelo.
    - Se aplica M_A a los Rechazados y se estima su P(B|X) y P(G|X)=1- P(B|X)
    - Cada R se duplica en la base de datos, un caso como B y otro como G, con ponderaciones respectivas P(B|X) y P(G|X)
    - Obtener un modelo del proceso de aprobación (como en Augmentation), que provea una estimacion de P(A|Y) siendo Y variables predictivas.
    - Combinar A y R con ponderaciones considerando la probabilidad de ser aprobado P(A|Y)
  - Imputacion por vecinos cercanos
    - Se utilizan técnicas de clustering para asignar los R de acuerdo a sus vecinos más cercanos, pudiendo estimar una P(B) de acuerdo a la frecuencia relativa de sus vecinos.
  - Muchos otros más (imputacion múltiple, Mixture Models, adaptaciones de la corrección de Heckman, aprendizaje semi-supervisado, etc)

Inferencias…resumiendo

BeSmart.
SMART BUSINESS SOLUTIONS

1. Las inferencias no pueden ser validadas directamente, si bien hay estudios basados en simulación orientados a seleccionar las mejores técnicas
2. Se deben hacer verificaciones tales como
   - La distribucion de score de los R queda a la izquierda de los A
   - La tasa de evento (inferida) de los R es mayor a la tasa de evento de los A
3. La única manera de conocer el comportamiento real consiste en definir una estrategia de testeo / control, como parte del proceso de originación, en la cual se aprueba un subset de los que serían rechazados.
4. Una aproximación a lo anterior es complementar con información de bureau de crédito (comportamiento con otros bancos que otorgaron crédito al Rechazado)
5. Se recomienda aplicar como mínimo, ante un sesgo no muy pronunciado, y cuando no se espera que haya en los datos distorsiones mayores (políticas que producen “cherry picking”), un método mínimo tal como parceling, de manera de reconstruir las estadísticas poblacionales.

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

BeSmart.
SMART BUSINESS SOLUTIONS

Caso 2: Detección del Hurto de Electricidad

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Detección del Hurto de Electricidad

BeSmart.
SMART BUSINESS SOLUTIONS

Hurto de electricidad: Manipulación de los medidores y/o otras acciones, acometidas por parte del suscriptor, a fin de lograr que sus registros sean inferiores a los que realmente deberían ser.

Métodos más habituales de adulteración:

- Bypass del medidor
- Intercambio de fase y neutro
- Empujar el vidrio para detener la rotación
- Insertar films, imanes o materiales similares para detener o dificultar la rotación
- Perforar agujeros en la rueda e insertar objetos para detener la rotación
- Romper los sellos y manipular los indicadores de consumos
- Daños en el sistema por someter el mecanismo a golpes
- Inclinar los medidores
- Uso de polvo o líquido viscoso en el disco giratorio
- Dañar la bobina mediante cortocicuito
- …

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Detección del Hurto de Electricidad

BeSmart.
SMART BUSINESS SOLUTIONS

- El consumo no registrado (CNR) produce pérdidas millonarias para las distribuidoras de electricidad.
- El hurto y otras Pérdidas No Técnicas constituyen un drenaje de recursos energéticos escasos.
- Las conexiones clandestinas impactan en la calidad del servicio y en la seguridad de las instalaciones.

- Para detectar el hurto, se llevan a cabo inspecciones de los medidores e instalaciones.
- Fuimos seleccionados por una de las mayores empresas distribuidoras locales para **proveer un sistema de priorización de las inspecciones basado en minería de datos**.

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Fuentes de datos analizadas

BeSmart.
SMART BUSINESS SOLUTIONS

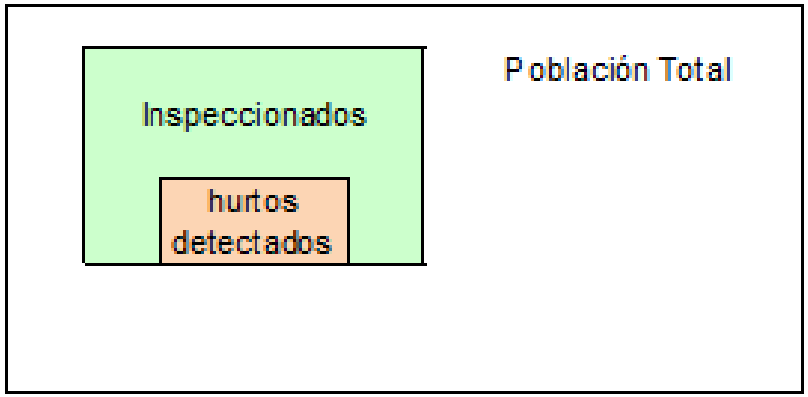
- **Historia de datos analizada: 7+ años**
- **2.5MM clientes**

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Donde está el sesgo de selección?

BeSmart.
SMART BUSINESS SOLUTIONS

- Procedimiento de priorización de las inspecciones no es al azar
- Es el que da origen a los datos!
- Priorización de inspecciones basada en reglas de experto, del estilo:
    - Inspeccionar si el consumo C cayó más de X% y $C>C_0$ (comparaciones interanuales y con el mes/bimestre anterior
    - Inspeccionar si lleva 3 lecturas con consumo en cero
    - Inspeccionar si lleva 4 lecturas con consumo bajo $C<C_0$
    - Etc
- Sólo conocemos si el cliente está hurtando, si ha sido inspeccionado.
- De la población no inspeccionada (en un determinado punto en el tiempo), no conocemos si está hurtando (es decir, no conocemos la variable target)



UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Medidas para mitigar el sesgo

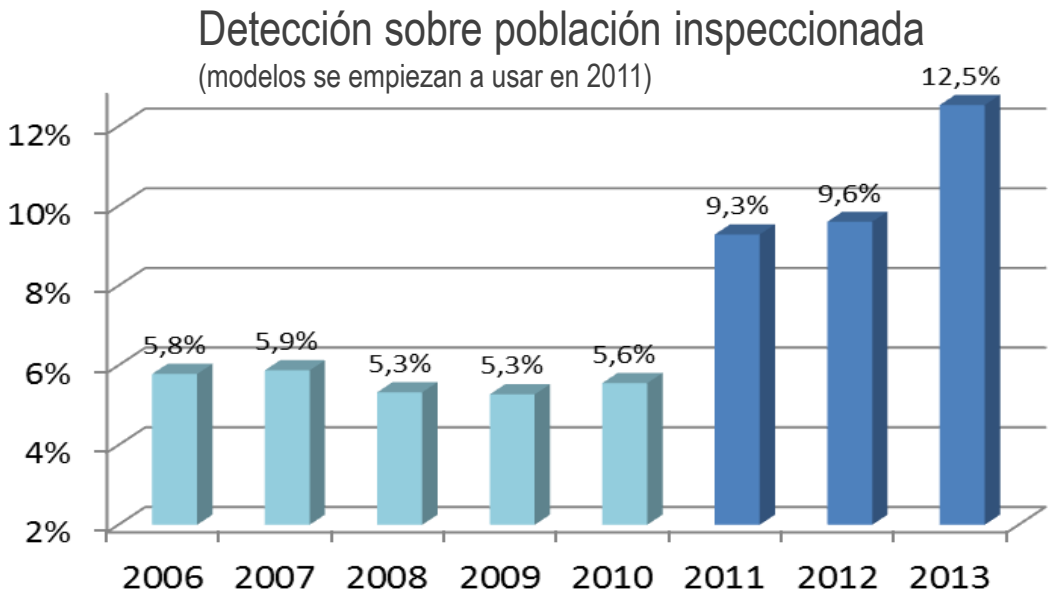
BeSmart.
SMART BUSINESS SOLUTIONS

- Se identificaron, dentro de los segmentos de interés (Residenciales por region, Tarifa General por región, Pequeñas Empresas, etc), cuales tenían un potencial sesgo mayor en funcion de:
  - La tasa de inspección (TI)
  - La tasa de hurto (TH)
  - La relación entre ambas tasas (1-TI)/TH
  - Segmentos más afectados: residenciales (menor tasa de inspección dado el costo que implicaría barrer una población de más de 2MM de clientes)
- Se compararon estadísticas de variables representativas en las poblaciones Inspeccionadas (IN) y No Inspeccionadas (NI).
- Se obtuvieron modelos para la población inspeccionada (probabilidad de Hurto P(H)), y modelos para la probabilidad de ser inspeccionado P(IN). Dentro de las técnicas de modelización analizadas, se priorizaron aquellas en las que el output tuvieran una traducción directa a probabilidades (logit, probit), cuando no hubiera pérdida sustancial de poder predictivo.
- Se aplicaron métodos de inferencia para la reconstrucción de las estadísticas poblacionales y para validar que la selección de variables fuera robusta.
- Se estimó la tasa de detección a obtener con diferentes puntos de corte (para la estrategia de uso del modelo), y esta fue validada durante los primeros meses de aplicación del modelo, junto con otros testeos.

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Resultados detección de hurto

- Proyecto pionero en la región
- La aplicación sistemática de los modelos para la priorización de inspecciones muestra un incremento sustancial en la detección de hurto:
  - ✓ Duplicación de la tasa de detección para importantes poblaciones, respecto del método previo, con un volumen de inspecciones comparable a la etapa previa.
  - ✓ Se obtuvo un rápido retorno a la inversión gracias al sustancial incremento en la recuperación
  - ✓ A medida que se avanzó en el uso se logró un aprendizaje continuo, ajustando algunos modelos.



Plan de proyecto adaptado

BeSmart.
SMART BUSINESS SOLUTIONS

Comprensión del Negocio y de los datos

- Reuniones de relevamiento
- **Análisis de procedimientos actuales de priorización de inspecciones**
- Identificación de fuentes de datos, historia disponible
- Segmentación de la población para determinar los modelos e indicadores a desarrollar

Acceso a los datos

- Definición del conjunto de informaciones históricas a ser exploradas durante el proceso de Minería de Datos.
- Procedimientos de anonimización
- Generación de pedidos a Sistemas .

Análisis de calidad, consistencia y preparación de datos

- Testeos rigurosos de calidad y consistencia de datos
- Comprensión adecuada de los mismos, codificaciones y relaciones.

Desarrollo de modelos

- Evaluación del sesgo de las muestras (por segmento de población)
- Aplicación de técnicas de Modelización por segmento
- Estudio de inferencia de población no inspeccionada
- Patrones de comportamiento son observados en muestras independientes en distintos puntos del tiempo.

Proceso mensual

- Actualización mensual de la información que forma parte de los modelos
- Puntajes e indicadores se calculan en forma mensual para la priorización de las inspecciones, y se ponen a disposición de las áreas usuarias.
- Monitoreo
- Ajustes y seguimiento con las áreas usuarias

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Discusión Final

BeSmart.
SMART BUSINESS SOLUTIONS

- Hemos analizado situaciones en las que el sesgo no puede ser ignorado.
- El objetivo principal de hoy ha sido desarrollar una intuición práctica acerca del tema.
- El caso de Scoring de originación de créditos (inferencia de rechazos) es conocido y ha sido estudiado en profundidad. Sin embargo en muchos casos sigue siendo ignorado.
- Hemos visto otro ejemplo, **detección del hurto de electricidad**, en que surge la necesidad de estudiar y mitigar el sesgo muestral para obtener modelos estables y estimaciones precisas (para diseñar estrategias).
- Creemos que en la medida en que más fuentes de datos e industrias se “someten” a la explotación de sus datos, los efectos de sesgos en los mismos continuarán apareciendo.
- Se puede adaptar un plan de proyecto que contemple la detección, evaluación y eventual corrección del sesgo.
- Se recomienda adaptar procesos para que incorporen estrategias “Champion Challenger” (grupos de testeo/control)

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

GRACIAS

fcastelpoggi@besmart.com.ar

www.besmart.com.ar

BeSmart.

SMART BUSINESS SOLUTIONS

Método de Parceling

- Evalúo tasa de Rechazos

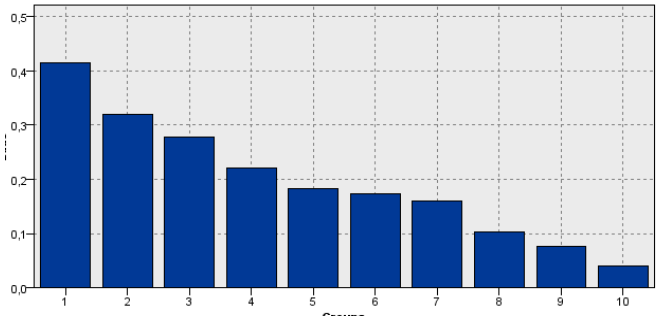
R	A	Tasa_Rech
2628	10162	0.205

- Tasa de Eventos (Aprobados)

B	G	Tasa_Evento
1966	8196	0.193

- Modelo sobre Aprobados:

SCORE_GB_Mean	Prob_B_Mean	logit_GB_Mean	Recuento_registros
803.367	0.196	1.589	10162

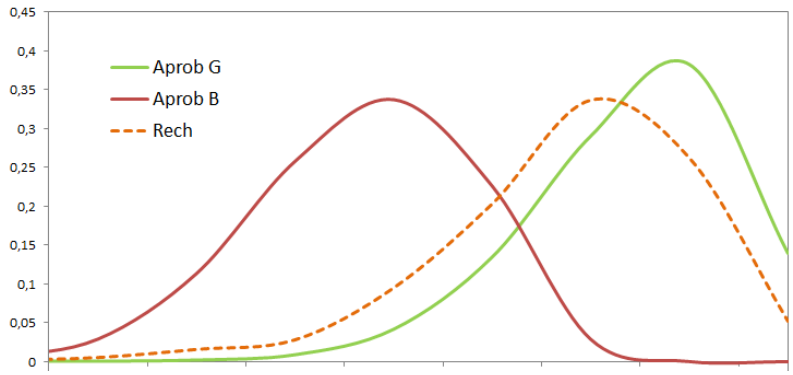


- Aplico sobre Rechadados:

SCORE_GB_Mean	Prob_B_Mean	logit_GB_Mean	Recuento_registros
774.368	0.225	1.418	2628

G	B	Tasa_Evento
2018	610	0.232

- Comparo distribuciones:



UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Assessing New Features for Online Advertisers

Diane Lambert
Google, Inc.
76 Ninth Ave
New York, NY 10011
dlambert@google.com

Daryl Pregibon
Google, Inc.
76 Ninth Ave
New York, NY 10011
daryl@google.com

ABSTRACT

Online search systems that display ads continually offer new features that advertisers can use to fine-tune and enhance their ad campaigns. An important question is whether a new feature actually helps advertisers. In an ideal world for statisticians, we would answer this question by running a statistically designed experiment. But that would require randomly choosing a set of advertisers and forcing them to use the feature, which is not realistic. Accordingly, in the real world, new features for advertisers are seldom evaluated with a traditional experimental protocol. Instead, customer service representatives select advertisers who are invited to be among the first to test a new feature (i.e., white-listed), and then each white-listed advertiser chooses whether or not to use the new feature. Neither the customer service representative nor the advertiser chooses at random.

This paper addresses the problem of drawing valid inferences from whitelist trials about the effects of new features on advertiser happiness. We are guided by three principles. First, statistical procedures for whitelist trials are likely to be applied in an automated way, so they should be robust to violations of modeling assumptions. Second, standard analysis tools should be preferred over custom-built ones, both

General Terms

Statistical Inference, Biased Sampling, Propensity Scores, Causal Modeling

1. INTRODUCTION

Randomized experiments are commonplace in the search engine (SE) industry. They are used to evaluate new ranking functions, changes to the user interface, and new algorithms for ad placement. These changes are typically tested on a sample of users that are chosen by randomly directing each query or cookie (depending on the study design) to the new conditions or the standard operating conditions.

Advertisers provide the revenue that allows search engines to provide free services to their users. SEs do their best to ensure that advertisers, like users, are happy by providing tools to manage and tune ads campaigns. These tools are continually improved and expanded according to changing business needs and advertiser feedback. When a new feature is introduced in the ads system front-end, it is often tested on a selected (i.e., white-listed) subset of advertisers before it is introduced to the entire advertiser base. Developers have limited control over which advertisers are white-listed,

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks. This work also suggests that, in contrast to prevailing assumptions, in-person interaction and nonverbal cues are not strictly necessary for emotional contagion, and that the observation of others' positive experiences constitutes a positive experience for people.

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are later seen by their friends via Facebook's "News Feed" product (8). Because people's friends frequently produce much more content than one person can view, the News Feed filters posts, stories, and activities undertaken by friends. News Feed is the primary manner by which people see content that friends share. Which content is shown or omitted in the News Feed is determined via a ranking algorithm that Facebook co develops and tests in the interest of showing viewers the they will find most relevant and engaging. One suc reported in this study: A test of whether posts with e content are more engaging.

The experiment manipulated the extent to which peo