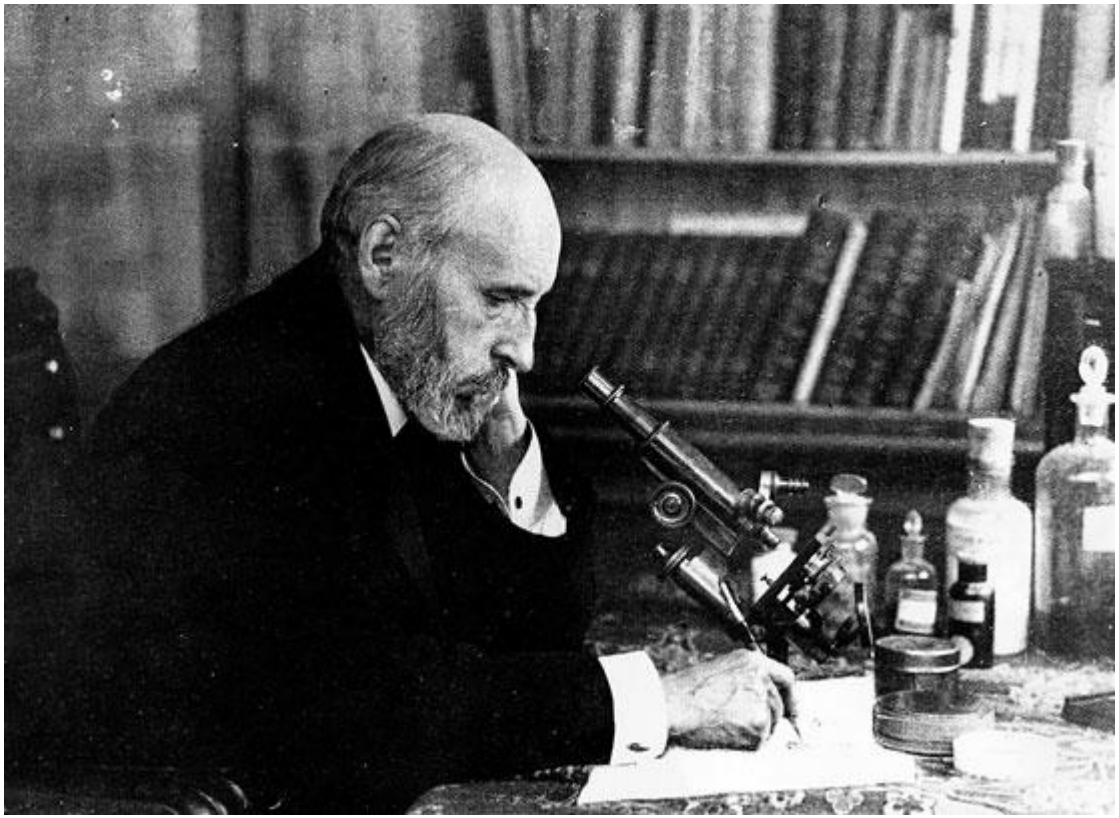
¿Factores de éxito para ganar un Nobel?



¿Factores de éxito para ganar un Nobel?

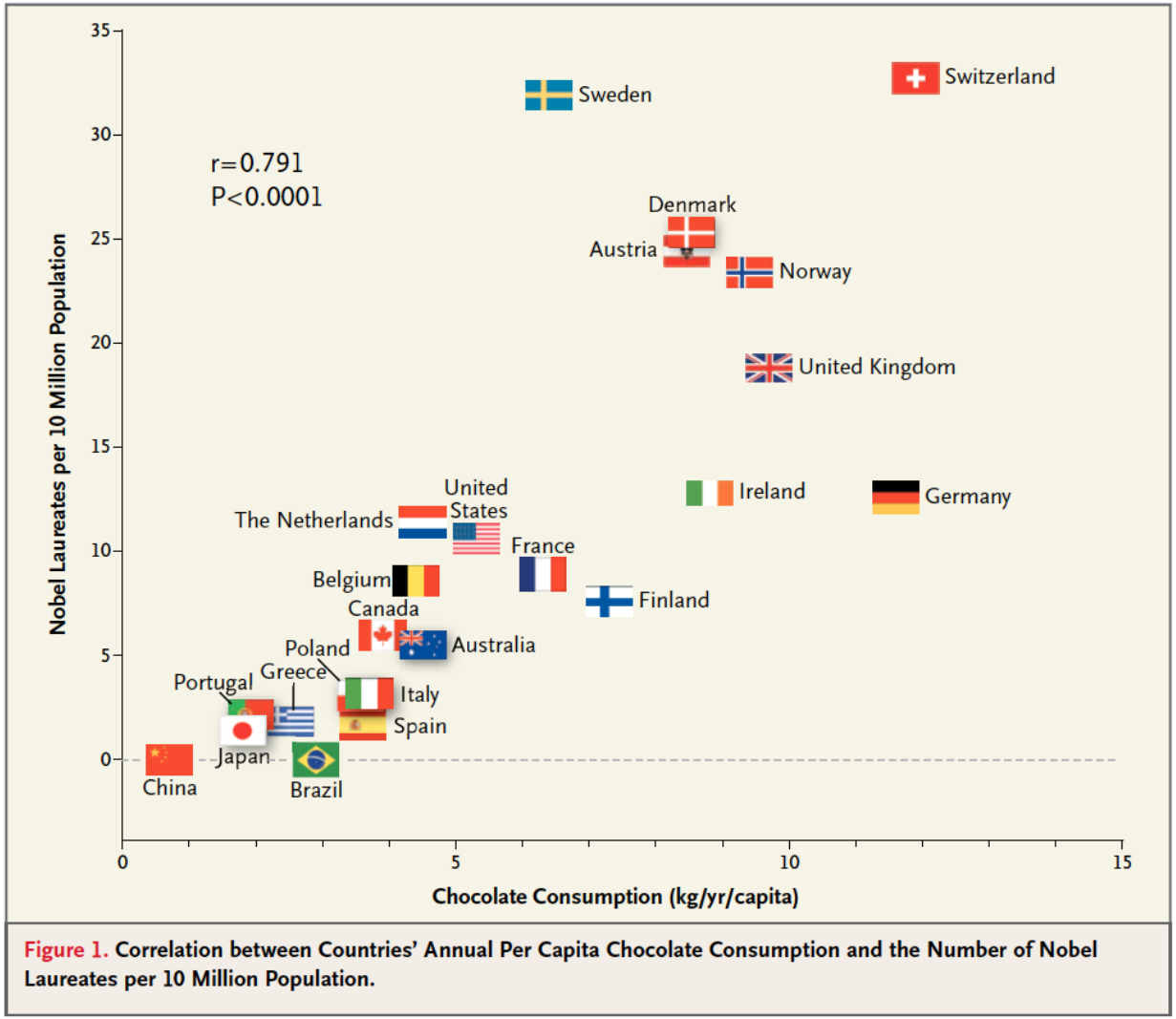
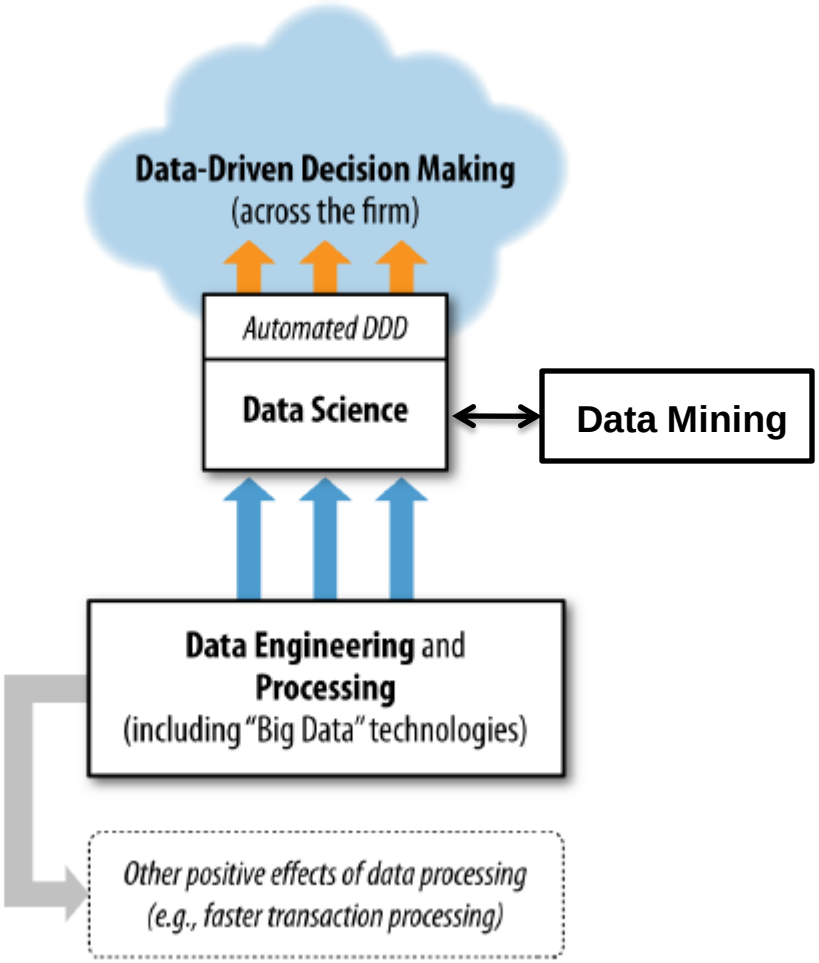
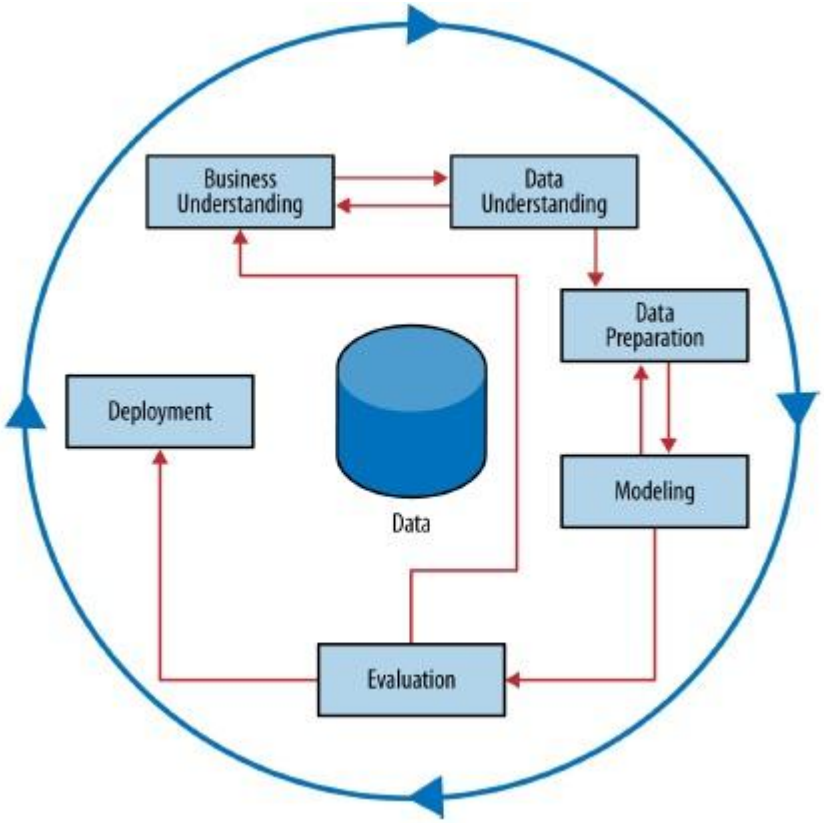


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Objetivo: Toma de decisiones basada en datos



Del Business Understanding al Data Mining



Problema: Campaña de Fundraising

For this case study, let's consider a real example of targeted marketing: targeting the best prospects for a charity mailing. Fundraising organizations (including those in universities) need to manage their budgets and the patience of their potential donors. In any given campaign segment, they would like to solicit from a "good" subset of the donors. This could be a very large subset for an inexpensive, infrequent campaign, or a smaller subset for a focused campaign that includes a not-so-inexpensive incentive package.

So let's get specific. A data miner might immediately think: we want to model the probability that each prospective customer, a prospective donor in this case, will respond to the offer. However, thinking carefully about the business problem we realize that in this case, the response can vary—some people might donate $100 while others might donate $1. We need to take this into account.

Objetivo: Diseñar una campaña para maximizar la captación de fondos, dónde hay que descontar a cada donación el coste de la solicitud.

Modelo de toma de decisiones

Expected Value Framework (Decisión basada en el valor esperado):

$$\text{Expected benefit of targeting} = p(R \mid \mathbf{x}) \cdot v_R(\mathbf{x}) + [1 - p(R \mid \mathbf{x})] \cdot v_{NR}(\mathbf{x})$$

donde:

P(R | x) es la probabilidad de que la persona x responda positivamente.

v_R (x) es el valor obtenido si la respuesta de x es positiva

v_{NR} (x) es el valor obtenido si la respuesta de x es negativa

v(x) = donacion(x) – coste(x)

Modelo de toma de decisiones

El objetivo consiste en diseñar una campaña tal que:

Expected benefit of targeting > 0

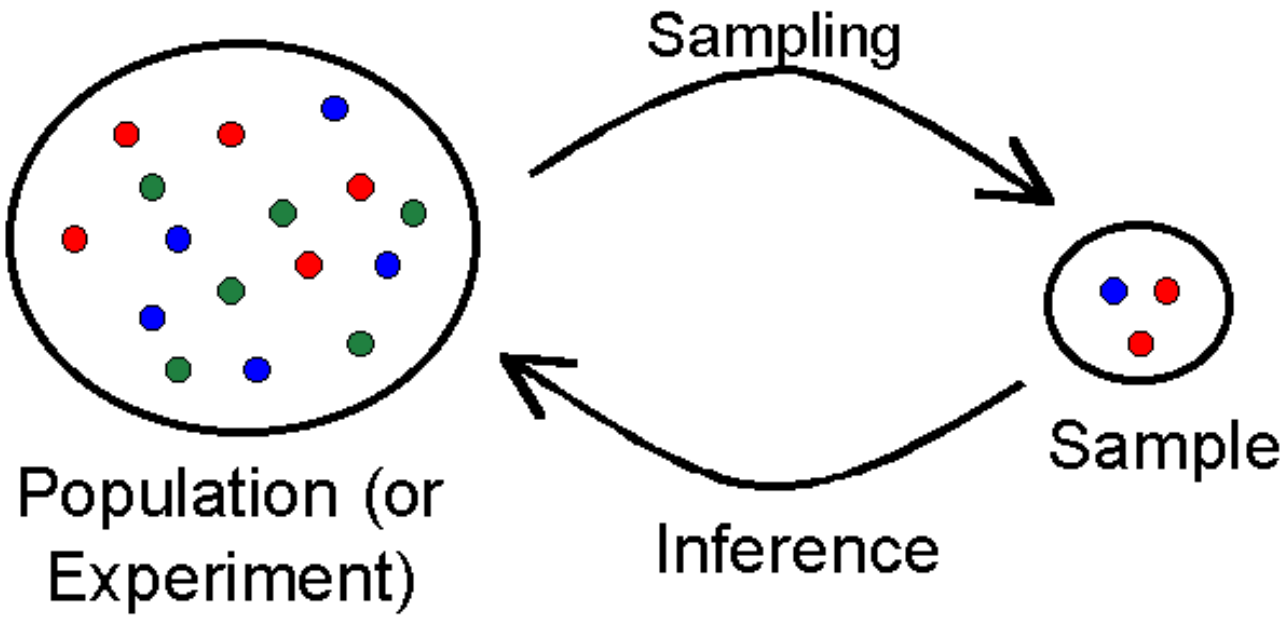
Lo que implica que el beneficio tiene que ser superior al coste**:**

$$
\begin{aligned}
p(R \mid \mathbf{x}) \cdot (d_R(\mathbf{x}) - c) + [1 - p(R \mid \mathbf{x})] \cdot (-c) &> 0 \\
p(R \mid \mathbf{x}) \cdot d_R(\mathbf{x}) - p(R \mid \mathbf{x}) \cdot c - c + p(R \mid \mathbf{x}) \cdot c &> 0 \\
p(R \mid \mathbf{x}) \cdot d_R(\mathbf{x}) &> c
\end{aligned}
$$

Ejercicio: ¿Cómo se resuelve el problema?

1. **¿El modelo representa correctamente el problema?**
2. **¿Qué supuestos asume el modelo?¿Qué supuestos no asume y sería correcto considerar?**
3. **¿Qué datos necesitamos para conocer las variables?**
4. **¿Cómo podemos obtener esos datos?**
5. **¿Cómo se puede evaluar el modelo antes de lanzar la campaña?**

El problema del Sesgo de la Muestra



Problema: Pérdida o deserción de clientes

Assume you just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the United States. They are having a major problem with customer retention in their wireless business. In the mid-Atlantic region, 20% of cell phone customers leave when their contracts expire, and it is getting increasingly difficult to acquire new customers. Since the cell phone market is now saturated, the huge growth in the wireless market has tapered off. Communications companies are now engaged in battles to attract each other's customers while retaining their own. Customers switching from one company to another is called *churn*, and it is expensive all around: one company must spend on incentives to attract a customer while another company loses revenue when the customer departs.

You have been called in to help understand the problem and to devise a solution. Attracting new customers is much more expensive than retaining existing ones, so a good deal of marketing budget is allocated to prevent churn. Marketing has already designed a special retention offer. Your task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to decide which customers should be offered the special retention deal prior to the expiration of their contracts.

Think carefully about what data you might use and how they would be used. Specifically, how should MegaTelCo choose a set of customers to receive their offer in order to best reduce churn for a particular incentive budget? Answering this question is much more complicated than it may seem initially. We will return to this problem repeatedly through the book, adding sophistication to our solution as we develop an understanding of the fundamental data science concepts.

Objetivo: Minimizar la probabilidad de deserción seleccionando un cto de usuarios a los que se les ofrecerá una oferta especial para que continúen.

Modelo I del problema de la deserción

Expected Value Framework (Decisión basada en el valor esperado):

$$\text{Expected benefit of targeting} = p(S \mid \mathbf{x}) \cdot v_S(\mathbf{x}) + [1 - p(S \mid \mathbf{x})] \cdot v_{NS}(\mathbf{x})$$

donde:

P(S | x) es la probabilidad de que la persona x se quede en la empresa.

v_S (x) es el valor obtenido si la persona x se queda

v_{NR} (x) es el valor obtenido si la persona x se va de la empresa

v(x) = cuota_anual(x) – oferta_prorrateada_anual(x)

Modelo I del problema de la deserción

¿Es correcto el modelo I?:

$$\text{Expected benefit of targeting} = p(S \mid \mathbf{x}) \cdot v_S(\mathbf{x}) + [1 - p(S \mid \mathbf{x})] \cdot v_{NS}(\mathbf{x})$$

¿Qué pasa con el valor de aquellos usuarios a los que no dirigimos la campaña?

¿El beneficio esperado es cero para aquellos usuarios
a los que no dirigimos la campaña?

¡NO!

Modelo II del problema de la deserción

Hay que considerar dos modelos, uno para el grupo al que dirigimos las ofertas, y otro para el que no**:**

$$EB_T(\mathbf{x}) = p(S \mid \mathbf{x}, T) \cdot (u_s(\mathbf{x}) - c) + [1 - p(S \mid \mathbf{x}, T)] \cdot (u_{NS}(\mathbf{x}) - c)$$

$$EB_{notT}(\mathbf{x}) = p(S \mid \mathbf{x}, notT) \cdot (u_s(\mathbf{x}) - c) + [1 - p(S \mid \mathbf{x}, notT)] \cdot (u_{NS}(\mathbf{x}) - c)$$

Y el valor de la campaña de ofertas dependerá de la diferencia**:**

$$VT = EB_T(\mathbf{x}) - EB_{notT}(\mathbf{x})$$

$$\begin{aligned} VT &= p(S \mid \mathbf{x}, T) \cdot u_S(\mathbf{x}) - p(S \mid \mathbf{x}, notT) \cdot u_S(\mathbf{x}) - c \\ &= [p(S \mid \mathbf{x}, T) - p(S \mid \mathbf{x}, notT)] \cdot u_S(\mathbf{x}) - c \\ &= \Delta(p) \cdot u_S(\mathbf{x}) - c \end{aligned}$$

Modelo II del problema de la deserción

¿Interpretación del Modelo II?**:**

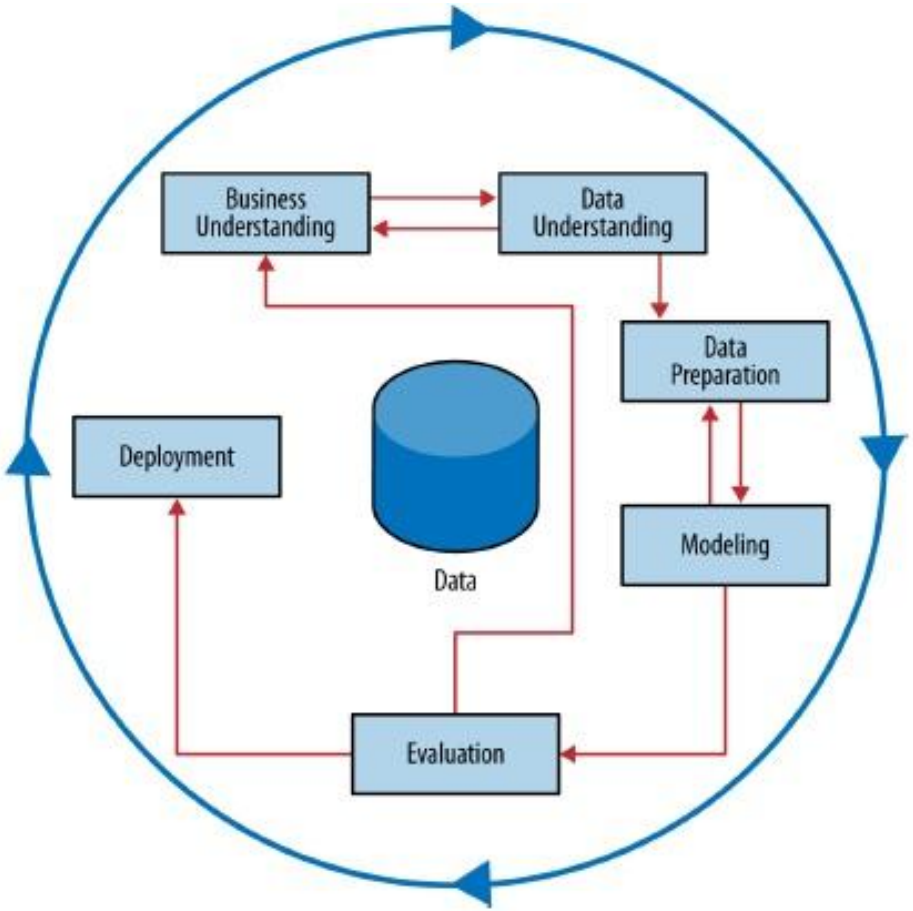
$$
\begin{aligned}
VT &= p(S \mid \mathbf{x}, T) \cdot u_S(\mathbf{x}) - p(S \mid \mathbf{x}, notT) \cdot u_S(\mathbf{x}) - c \\
&= [p(S \mid \mathbf{x}, T) - p(S \mid \mathbf{x}, notT)] \cdot u_S(\mathbf{x}) - c \\
&= \Delta(p) \cdot u_S(\mathbf{x}) - c
\end{aligned}
$$

1. El valor de la campaña dependerá del incremento de probabilidad de que x se quede si le hacemos la oferta.
2. Si la campaña no tiene el impacto esperado, la probabilidad no variará y sólo tendremo coste sin beneficio.

Ejercicio: ¿Cómo estimamos las variables?

1. **¿Cómo se calcula P(S | x, notT)?**
2. **¿ Cómo se calcula P(S | x, T)?**

Guía para la creación/evaluación de proyectos



Guía para la creación/evaluación de proyectos

1. **Entender el problema. Preguntas:**

    - ¿En qué consiste el problema de forma precisa?
    - ¿Qué entidades incluye el problema?
    - ¿Cuáles son las variables/atributos relevantes del problema?

2. **Entender los datos que necesitamos. Preguntas:**

    - ¿Tengo los datos?
    - ¿Puedo acceder a todas las variables?
    - ¿Cómo puedo obtener los datos?

3. **Preparar los datos. Preguntas**

    - ¿En qué formato tengo los datos?
    - ¿Tengo muchas variables? ¿Son todas necesarias?
    - ¿Tengo muchas/pocas observaciones?¿Cuantas necesito?

Guía para la creación/evaluación de proyectos

4. Crear un modelo. Preguntas:

- ¿Qué modelo(s) es(son) el(los) más apropiado(s)?
- Si hay varios posibles, ¿cuál escoger?
- ¿Tengo datos para estimar las variables del modelo?
- ¿Cumple el modelo aspectos no funcionales (generalización, escalabilidad, etc…)?

5. Evaluar el modelo. Preguntas:

- ¿Plan para evaluar el modelo?
- ¿Hay datos para evaluar el modelo?
- ¿Son apropiados los criterios (métricas) para evaluar el modelo?

Consejos para el Científico de Datos

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1. **“No tengo los datos para aplicar el modelo”. Preguntas:**

    - ¿Puedo calcular las variables actuales a través de otras que sí puedo observar?
    - ¿Puedo representar el problema de otro modo, de tal forma que dependa de otras variables?

2. **“El modelo es muy complejo, no sé cómo programarlo”. Preguntas:**

    - ¿Existe algún modelo más sencillo que describa el problema?
    - ¿La complejidad procede de modelar casos particulares que son improbables?

3. **“Tengo el modelo perfecto, pero mi muestra no se ajusta a los supuestos que asume el modelo”**

    - ¿Los supuestos del modelo son correctos?
    - ¿Es posible obtener nuevos datos con los supuestos correctos?

Ejercicio I: Churn reduction!

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Scenario and Proposal

You've landed a great job with Green Giant Consulting (GGC), managing an analytical team that is just building up its data science skill set. GGC is proposing a data science project with TelCo, the nation's second-largest provider of wireless communication services, to help address their problem of customer churn. Your team of analysts has produced the following proposal, and you are reviewing it prior to presenting the proposed plan to TelCo. Do you find any flaws with the plan? Do you have any suggestions for how to improve it?

Churn Reduction via Targeted Incentives — A GGC Proposal

We propose that TelCo test its ability to control its customer churn via an analysis of churn prediction. The key idea is that TelCo can use data on customer behavior to predict when customers will leave, and then can target these customers with special incentives to remain with TelCo. We propose the following modeling problem, which can be carried out using data already in TelCo's possession.

We will model the probability that a customer will (or will not) leave within 90 days of contract expiration, with the understanding that there is a separate problem of retaining customers who are continuing their service month-to-month, long after contract expiration. We believe that predicting churn in this 90-day window is an appropriate starting point, and the lessons learned may apply to other churn-prediction cases as well. The

Ejercicio I: Churn reduction!

model will be built on a database of historical cases of customers who have left the company. Churn probability will be predicted based on data 45 days prior to contract expiration, in order for TelCo to have sufficient lead time to affect customer behavior with an incentive offer. We will model churn probability by building an ensemble of trees (random forest) model, which is known to have high accuracy for a wide variety of estimation problems.

We estimate that we will be able to identify 70% of the customers who will leave within the 90-day time window. We will verify this by running the model on the database to verify that indeed the model can reach this level of accuracy. Through interactions with TelCo stakeholders, we understand that it is very important that the V.P. of Customer Retention sign off on any new customer retention procedures, and she has indicated that she will base her decision on her own assessment that the procedure used for identifying customers makes sense and on the opinions about the procedure from selected firm experts in customer retention. Therefore, we will give the V.P. and the experts access to the model, so that they can verify that it will operate effectively and appropriately. We propose that every week, the model be run to estimate the probabilities of churn of the customers whose contracts expire in 45 days (give or take a week). The customers will be ranked based on these probabilities, and the top N will be selected to receive the current incentive, with N based on the cost of the incentive and the weekly retention budget.

Ejercicio II: Whiz-Bang Widget

Your company has an installed user base of 900,000 current users of your Whiz-bang® widget. You now have developed Whiz-bang® 2.0, which has substantially lower operating costs than the original. Ideally, you would like to convert ("migrate") your entire user base over to version 2.0; however, using 2.0 requires that users master the new interface, and there is a serious risk that in attempting to do so, the customers will become frustrated and not convert, become less satistified with the company, or in the worst case, switch to your competitor's popular Boppo® widget. Marketing has designed a brand-new migration incentive plan, which will cost $250 per selected customer. There is no guarantee that a customer will choose to migrate even if she takes this incentive.

An external firm, Big Red Consulting, is proposing a plan to target customers carefully for Whiz-bang® 2.0, and given your demonstrated fluency with the fundamentals of data science, you are called in to help assess Big Red's proposal. Do Big Red's choices seem correct?

Ejercicio II: Whiz-Bang Widget

Targeted Whiz-bang Customer Migration—prepared by Big Red Consulting, Inc.

We will develop a predictive model using modern data-mining technology. As discussed in our last meeting, we assume a budget of $5,000,000 for this phase of customer migration; adjusting the plan for other budgets is straightforward. Thus we can target 20,000 customers under this budget. Here is how we will select those customers:

We will use data to build a model of whether or not a customer will migrate given the incentive. The dataset will comprise a set of attributes of customers, such as the number and type of prior customer service interactions, level of usage of the widget, location of the customer, estimated technical sophistication, tenure with the firm, and other loyalty indicators, such as number of other firm products and services in use. The target will be whether or not the customer will migrate to the new widget if he/she is given the incentive. Using these data, we will build a linear regression to estimate the target variable. The model will be evaluated based on its accuracy on these data; in particular, we want to ensure that the accuracy is substantially greater than if we targeted randomly.

To use the model: for each customer we will apply the regression model to estimate the target variable. If the estimate is greater than 0.5, we will predict that the customer will migrate; otherwise, we will say the customer will not migrate. We then will select at random 20,000 customers from those predicted to migrate, and these 20,000 will be the recommended targets.

Data Science y la Estrategia de Negocio

Eduardo M. Sánchez Vila

eduardo.sanchez.vila@usc.es

CITIUS

Grupo de Sistemas Inteligentes

Universidad de Santiago de Compostela