

## Modelos de clasificación

- Previo: heurística vs conclusiones estadísticas
1. Algoritmos: regresión logística, árboles, k-vecinos, ensembles y Naive-Bayes
  2. Métricas específicas de clasificación
  3. Caso en Big ML de clasificación binaria
  4. Clasificación múltiple
  5. Caso en Big ML de clasificación múltiple

## Heurística vs conclusiones estadísticas

La ciencia de datos es una disciplina que engloba tantos conocimientos de áreas distintas que sería muy complejo aplicar todo de forma rigurosa con una fiabilidad alta.

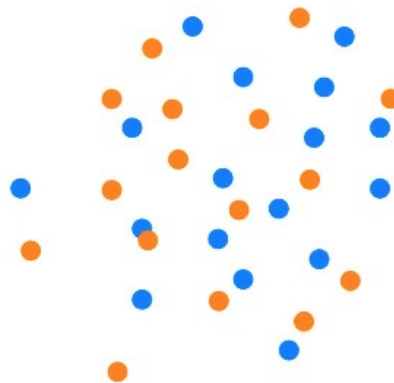
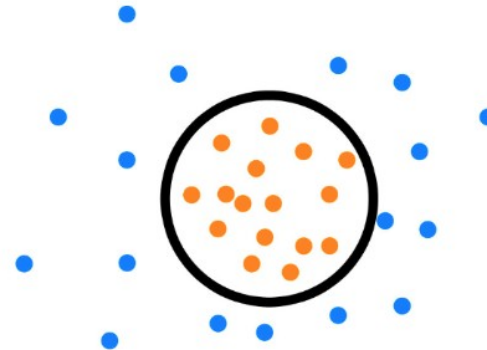
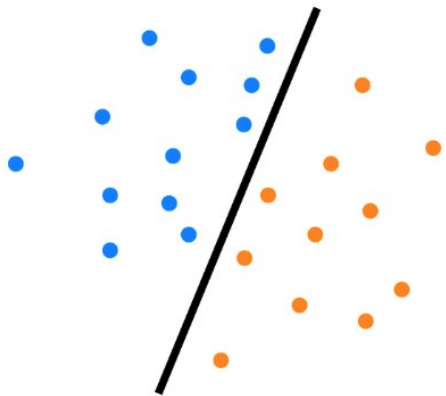
Es por esto que en general, incluso en grandes empresas que explotan muchos datos como pueden ser los vendedores de seguros, se aplican los modelos guiándose por la mera **heurística (dar una solución empírica que aproxime el problema y funcione en la práctica)** que en la **toma de decisiones basada en criterios estadísticos como intervalos de confianza y contrastes de hipótesis**.

Esto es así puesto que la necesidades de tiempo, desarrollo y aparataje matemático y estadístico en general sobrepasan las capacidades de análisis en la práctica.

No significa esto que se ignoren cuando se puedan usar, sino que se prefiere en muchas ocasiones la intuición y experiencia sobre el rigor.

## ML supervisado: clasificación

Los casos de clasificación son los más abundantes los entornos de negocio, particularmente la clasificación binaria 0 - 1



## ML supervisado: clasificación binaria

En el problema de clasificación binaria tenemos una entrada y queremos etiquetarla como 0 ó 1.

Los modelos de clasificación aportan una puntuación  $p$  entre 0 y 1 donde más próximo a 0 indica más probabilidad de etiquetado 0 y más próximo a 1 más probabilidad de etiquetado 1.

Generalmente, estableceremos un umbral  $U$  de modo que

- Predice 1 si la puntuación de predicción es mayor que  $U$
- Predice 0 si la puntuación de predicción es menor ó igual que  $U$

Para algunos modelos como la regresión logística, esta cantidad  $p$  de salida indica la **probabilidad de ser de la clase 1**. Las salidas de otros modelos son simples puntuaciones.

## Regresión logística

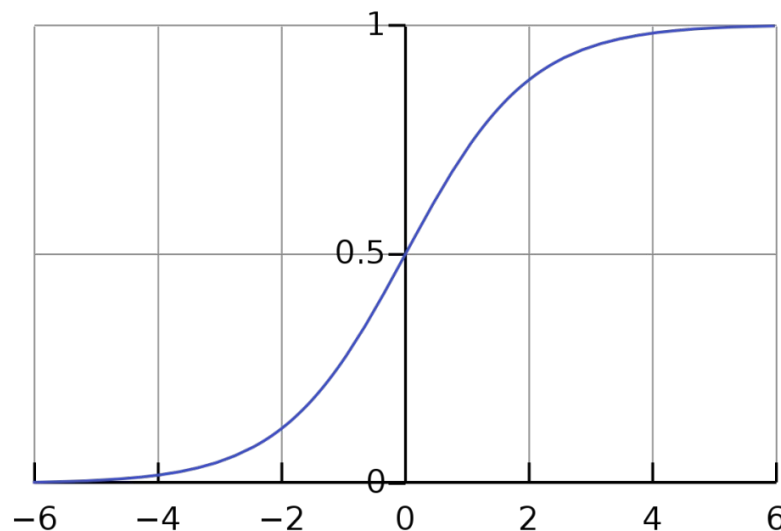
En la regresión logística modelamos el *logit*

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$$

linealmente respecto a los predictores

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

La regresión logística no es más que una regresión lineal cuya salida se ha transformado usando la **función sigmoide**, una de las funciones más importantes en ML



## Regresión logística: *log odds ratio*

Los resultados de los coeficientes de una regresión logística tienen interpretabilidad en forma de comparación de probabilidades

Suponemos que la probabilidad de éxito de un suceso es  $p = 0.8$ . Entonces el odd ratio es

$$\frac{p}{1-p} = \frac{0.8}{1-0.8} = 4$$

Esto significa que hay 4 veces más probabilidad de éxito que de fracaso, ya que el *odd ratio* es el cociente entre la probabilidad de éxito (1) y la de fracaso (0).

*Ejemplo: en una regresión logística donde se explica la "probabilidad de comprar un producto respecto al número de hijos", suponemos que el coeficiente del número de hijos es 0.1. Entonces el aumento de probabilidad de comprar el producto por cada hijo es de*

$$e^{\beta} = e^{0.1} = 1.105171$$

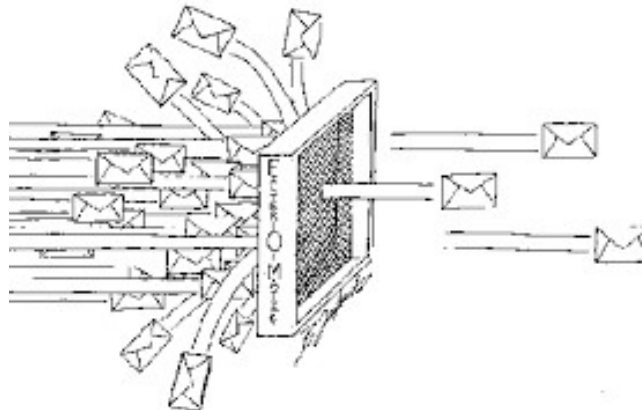
*Esto es, por cada hijo la probabilidad de que compre el producto aumenta en un 10.51%.*

## ML supervisado en clasificación: otros modelos

Así como en supervisado de regresión, se pueden aplicar también en clasificación los modelos:

- Árboles de decisión
- K - vecinos
- Modelos *ensemble*: RandomForest y Gradient Boosting

Además, para clasificación en el caso de tener muchísimos predictores, como puede ser el caso de un filtro de *spam*, se tienen modelos estilo Naive-Bayes. Estos modelos no requieren entrenamiento, ya que son meros cálculos de estimación realizados sobre los datos que aportan al final una probabilidad. Por lo general sólo se usan si hay muchos predictores y el problema es inabarcable, ya que dan rendimientos muy inferiores al resto.



## Métricas en clasificación y matriz de confusión

	Predicciones	
	Positivo	Negativo
Positivo en la realidad	TP	FN
Negativo en la realidad	FP	TN

Las descripciones de las componentes de la matriz de confusión:

- **TP**: cantidad de predicciones positivas que son realmente positivas
- **FP**: cantidad de predicciones positivas que son realmente negativas
- **TN**: cantidad de predicciones negativas que son realmente negativas
- **FN**: cantidad de predicciones negativas que son realmente positivas

- **TPR**: tasa de verdaderos positivos

$$TPR = TP/P = TP/(TP + FN)$$

- **FPR**: tasa de falsos positivos

$$FPR = FP/N = FP/(FP + TN)$$



## Métricas en clasificación y matriz de confusión

- **Accuracy:** precisión global a través de todas las clases.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** capacidad de acierto dentro de la clase predicha como positiva.

$$\frac{TP}{TP + FP}$$

- **Recall (igual a TPR):** capacidad de acierto dentro de la clase positiva.

$$\frac{TP}{TP + FN}$$

- **F1-score:** media armónica entre precision y recall. Es una métrica equilibrada entre tener pocos falsos positivos y pocos falsos negativos.

$$2 \frac{Precision * Recall}{Precision + Recall}$$

- **Área bajo la curva ROC:** valor entre 0.5 (clasificador aleatorio) y 1 (clasificador perfecto), que permite ajustar la TPR y la FPR

## Más métricas de clasificación

- **Correlación de Mathews ó coeficiente Phi:** está basado en la distancia usando el estadístico Chi Cuadrado normalizado entre histogramas. -1 indica modelo inverso, 0 es un modelo aleatorio y 1 un modelo perfecto.
- **Lift:** ratio que representa la ganancia de usar el modelo respecto a no usar ninguno (*random guessing*)
- **Estadístico K-S:** es el máximo valor de la diferencia entre TPR y FPR tomado entre todos los umbrales posibles
- **Tau de Kendall:** indicador entre -1 y 1. Más próximo a 1 indica un mejor modelo
- **Rho de Spearman:** indicador entre -1 y 1. Más próximo a 1 indica un mejor modelo. Valores negativos indican que el modelo invertido es mejor...

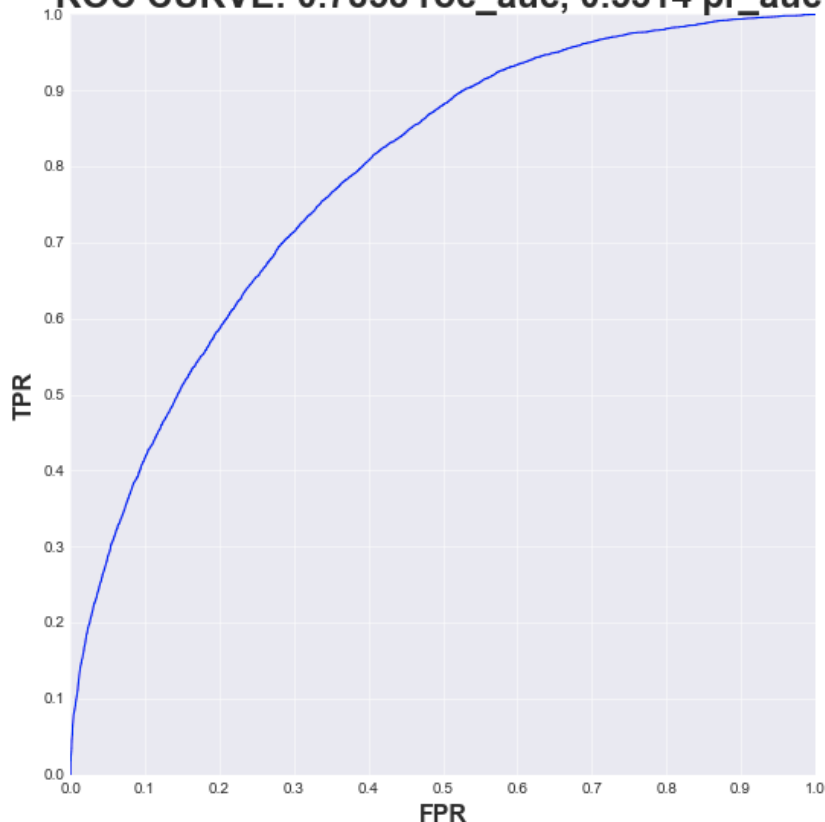
## Métricas en clasificación: curva ROC y umbrales

Al elegir el modelo con mayor **área bajo la curva ROC (AUC ROC)**, estamos también maximizando el potencial beneficio que aportará un modelo.

En función del tipo de problema son razonables distintos valores de ROC AUC.

Se elige un umbral que aporte FPR y TPR que maximice el ROI de la aplicación del modelo de clasificación. Cada FP tiene un coste y cada TP un beneficio.

**ROC CURVE: 0.7858 roc\_auc, 0.5314 pr\_auc**



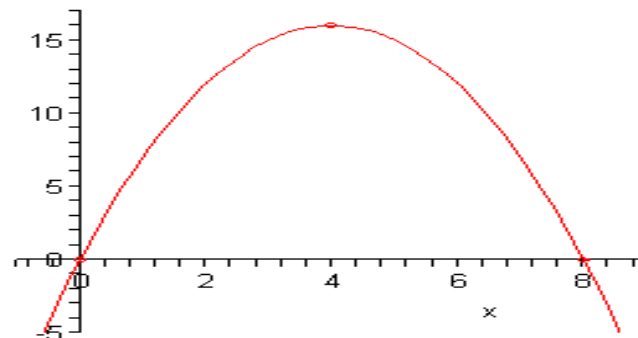
FPR (%)	TPR (%)
5	30
10	42
15	52
20	60
30	72

## Métricas en clasificación: maximización de ROI y selección de umbral

- **PR:** ratio de positivos. Es la proporción de positivos reales en los datos
- **NR:** ratio de negativos. Es la proporción de negativos reales en los datos
- **TPR:** la proporción de verdaderos positivos que da un predictor una vez seleccionado un umbral de decisión U
- **FPR:** la proporción de falsos positivos que da un predictor una vez seleccionado un umbral de decisión U
- **W:** ganancia estimada según las acciones tomadas ante un verdadero positivo
- **L:** pérdida estimada según las acciones tomadas ante un falso positivo

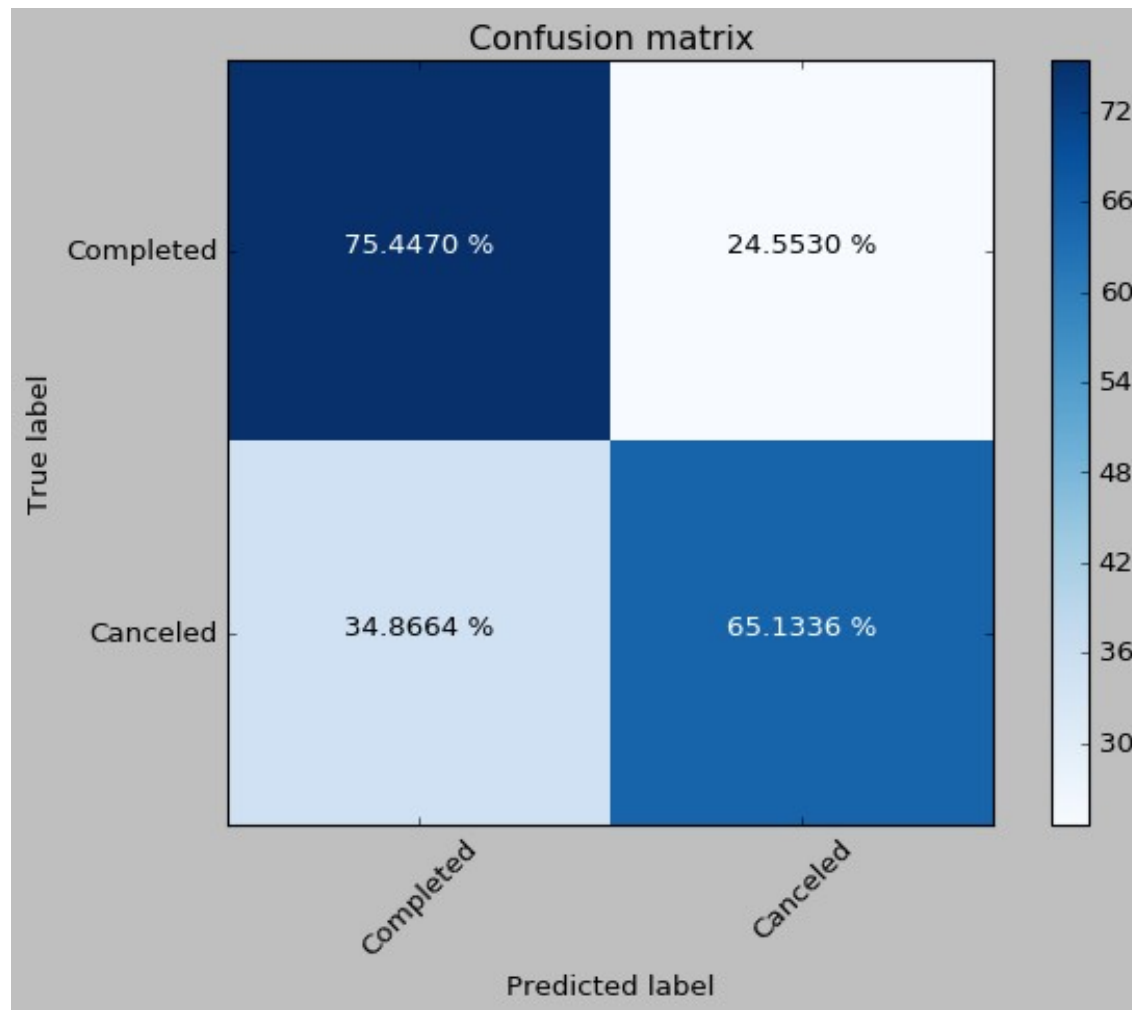
Elegiremos U de modo que los valores de TPR y FPR (que dependen de U) maximicen la expresión de beneficio:

$$\text{Beneficio}(U) = w \cdot PR \cdot TPR - l \cdot NR \cdot FPR$$



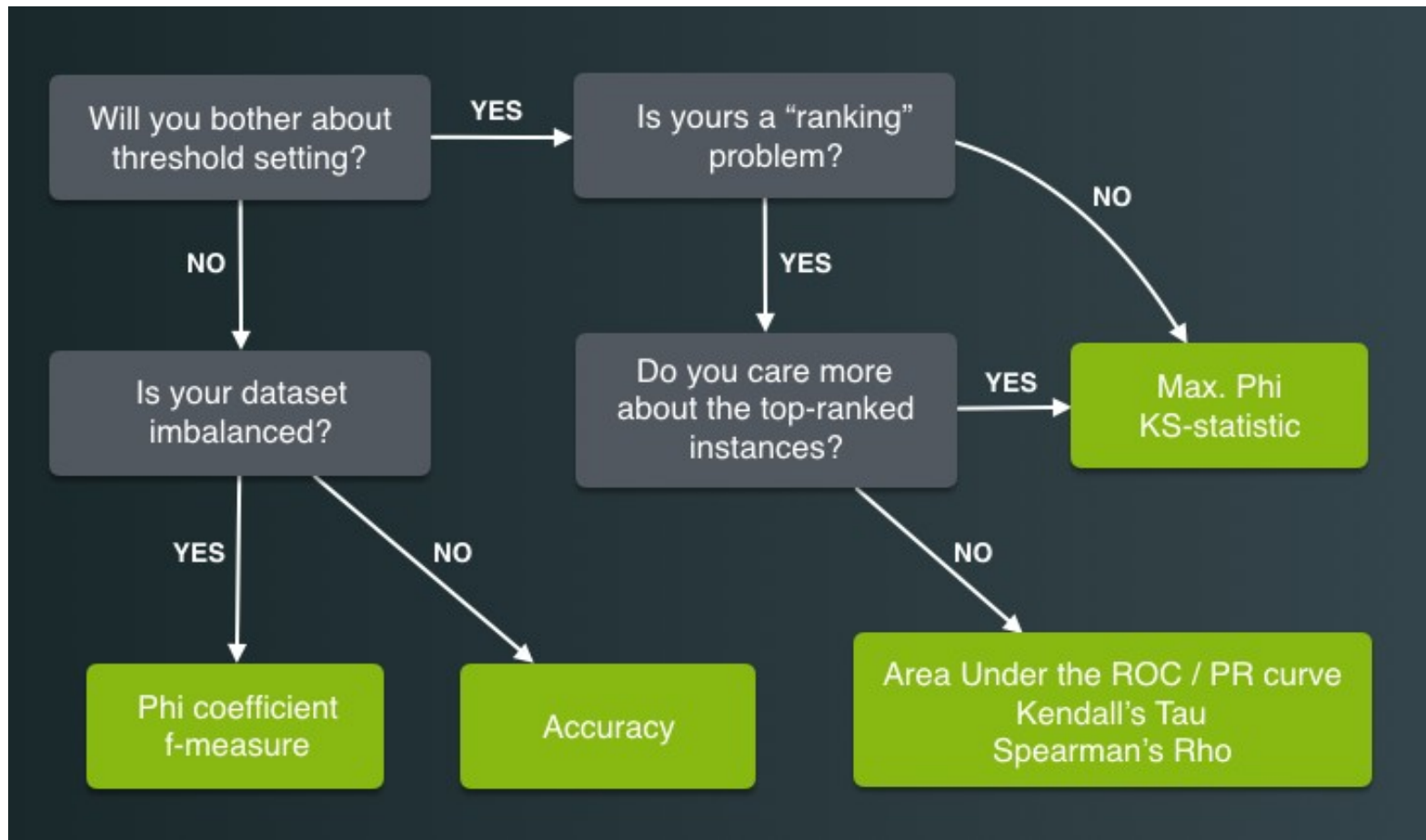
## Métricas en clasificación curva ROC y umbrales

Tras elegir un umbral en las puntuaciones del modelo clasificador, damos la matriz de confusión esperada:



## Métricas en clasificación curva ROC y umbrales

Selección de métricas de clasificación para elegir modelo:



## Aplicaciones de modelos de clasificación

### **Clasificación binaria:**

- Modelos de fraude
- Modelos de fuga
- Modelos de concesión de crédito
- Modelos de probabilidad de compra de un producto
- Modelos de *out-of-stock*
- Modelos de cancelación
- Modelos de éxito de un *trade*
- Modelo de análisis de sentimiento

### **Clasificación múltiple:**

- Modelo de clasificación de imágenes ó vídeos
- Modelo de clasificación de *topics* de documentos
- Modelo de clasificación de tipo de cliente respecto a productos

## Los procesos de la ciencia de datos

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<b>Determine Business Objectives</b> Background Business Objectives Business Success Criteria	<b>Collect Initial Data</b> Initial Data Collection Report  <b>Describe Data</b> Data Description Report  <b>Explore Data</b> Data Exploration Report  <b>Verify Data Quality</b> Data Quality Report	<b>Select Data</b> Rationale for Inclusion/Exclusion  <b>Clean Data</b> Data Cleaning Report  <b>Construct Data</b> Derived Attributes Generated Records  <b>Integrate Data</b> Merged Data  <b>Format Data</b> Reformatted Data  Dataset Dataset Description	<b>Select Modeling Techniques</b> Modeling Technique Modeling Assumptions  <b>Generate Test Design</b> Test Design  <b>Build Model</b> Parameter Settings Models Model Descriptions  <b>Assess Model</b> Model Assessment Revised Parameter Settings	<b>Evaluate Results</b> Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models  <b>Review Process</b> Review of Process  <b>Determine Next Steps</b> List of Possible Actions Decision	<b>Plan Deployment</b> Deployment Plan  <b>Plan Monitoring and Maintenance</b> Monitoring and Maintenance Plan  <b>Produce Final Report</b> Final Report Final Presentation  <b>Review Project</b> Experience Documentation
<b>Assess Situation</b> Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits					
<b>Determine Data Mining Goals</b> Data Mining Goals Data Mining Success Criteria					
<b>Produce Project Plan</b> Project Plan Initial Assessment of Tools and Techniques					

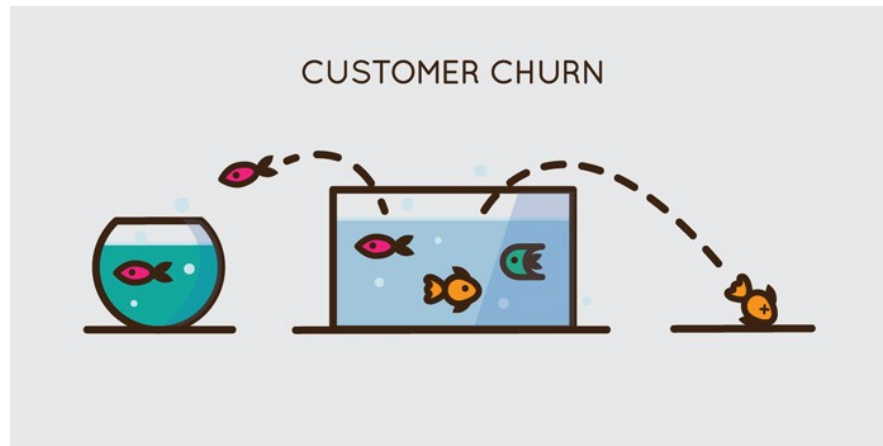


## Caso en Big ML de clasificación binaria: *Telecom churn*

Los modelos de fuga intentan captar las circunstancias que se verifican en un cliente que va a abandonar una compañía con antelación.

*¿un cliente seleccionado se va a dar de baja en los siguientes X meses?*

Esto permite tomar acciones como ofrecer promociones con permanencia que fidelicen al cliente durante un período.



Son mucho más costosas las campañas de captación que las medidas para fidelizar, por tanto los modelos de fuga son vitales hoy en día en mercados fuertemente competitivos como el de las telecomunicaciones.

## Caso en Big ML de clasificación binaria: *Telecom churn - background*

### Dimensión del problema

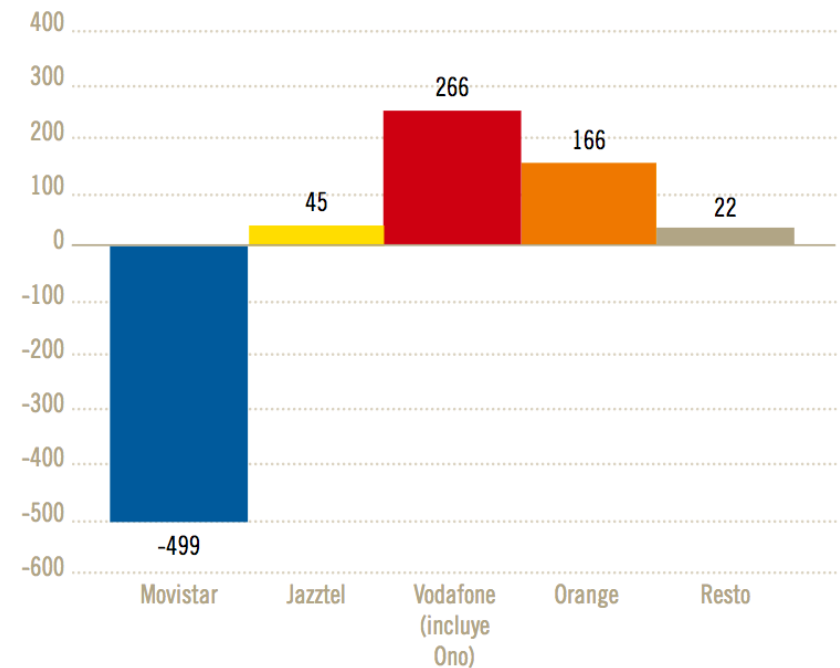
En el 2014 el *churn rate* fue del 29,1%, un de los más altos en el histórico de Europa.

**Portabilidad de líneas fijas (miles de líneas)**



Fuente: CNMC

**Saldo neto de portabilidad por operador en 2014 (miles de líneas)**



## Caso en Big ML de clasificación binaria: *Telecom churn - background*

### Documentación sobre estudios previos del problema

Intentamos aprender todo lo que se ha hecho para atacar este problema en la comunidad de la ciencia de datos.

Esto nos permite partir del **estado del arte**, y como consecuencia tendremos más capacidad de conseguir buenos rendimientos

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). **Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry**. *Telecommunications policy*, 30(10), 552-568.

Hypothesis test results: customer churn determinants

Dependent variable	Independent variable	Hypothesis	Result
Customer churn	Call drop rate	H1a	Accept
	Call failure rate	H1b	Reject
	Number of complaints	H1c	Accept
	Loyalty points	H2a	Accept
	Membership card	H2b	Reject
	Billed amounts	H3a	Accept
	Unpaid balances	H3b	Reject
	Number of unpaid monthly bills	H3c	Reject
	Customer status	H4	Accept

## Caso en Big ML de clasificación binaria: *Telecom churn - background*

### Predictores en el dataset

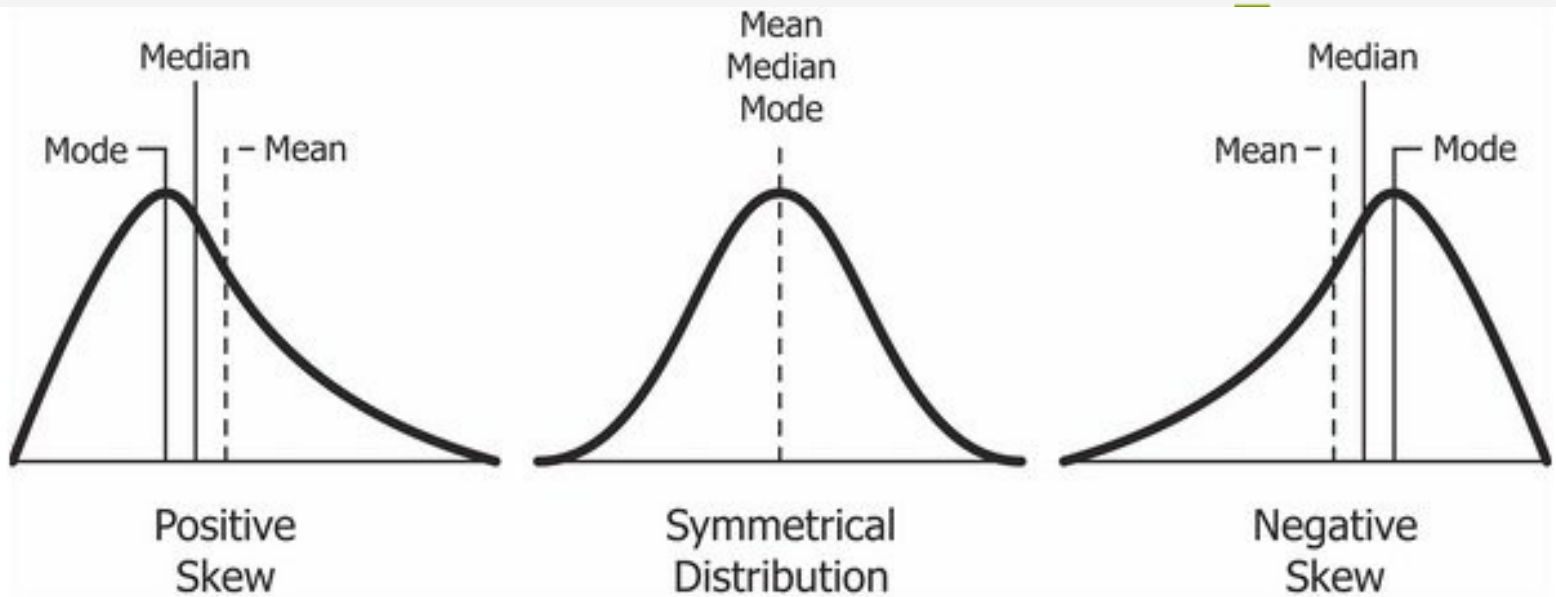
Analizamos el conjunto de predictores disponibles y buscamos ó planteamos potenciales predictores interesantes.

Los predictores que disponemos son de las siguientes categorías:

- **Características demográficas:** edad, sexo, ciudad de residencia, etc.
- **Características de soporte:** interacción del usuario con el servicio de asistencia: número de llamadas, cuestiones planteadas, valoración de su satisfacción.
- **Características de uso:** uso que hace el cliente del servicio: número de interacciones con el servicio, planes contratados, gasto mensual.
- **Características adicionales o de contexto:** otro tipo de información útil para la predicción (ej, antigüedad del cliente).

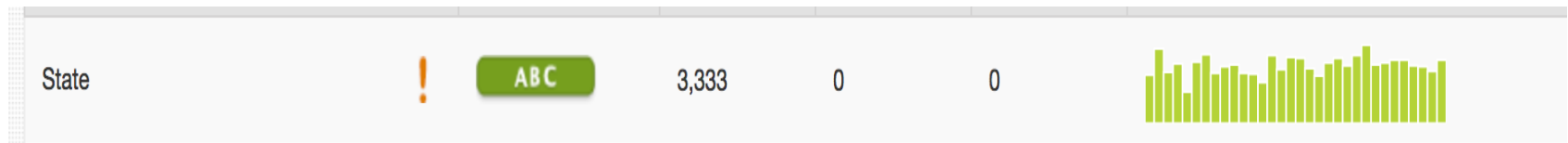
## Distribuciones en Big ML

- Para distribuciones no sesgadas (**skewed**), la media puede ser una buena indicación de un valor típico y ser menos sensible a **outliers**, pero solo en esos casos.
- Un recurso para eliminar ese sesgo es **tomar el logaritmo** de la variable
- La mediana es más robusta en otros casos.



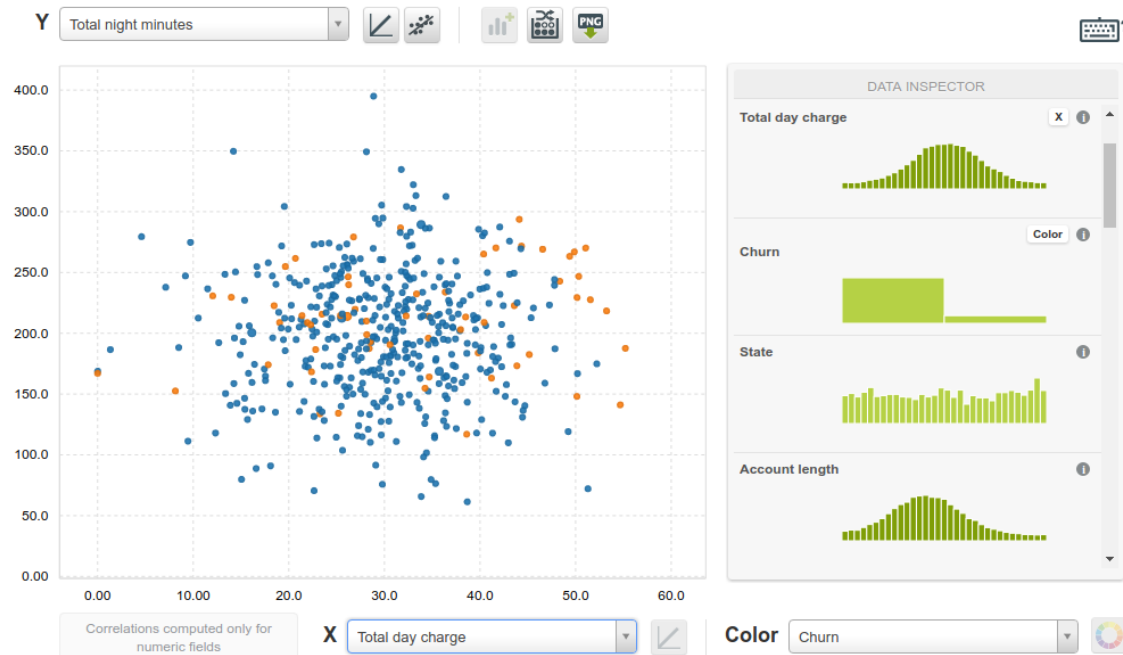
## Distribuciones en Big ML

- BigML descarta automáticamente variables basadas en su distribución.
- Esto puede o no ser correcto para lo que buscamos, especialmente en el caso de variables nominales.



## Caso en Big ML de clasificación binaria: *Telecom churn*

- Vemos la web <http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>
- Cargamos el dataset en el proyecto ML supervisado clasificación
- Exploramos el dataset creando imágenes como la de abajo y buscando separación
- ¿Qué variables tienen mayor capacidad predictiva?
- Observa la variable *churn*. Es una clase desequilibrada
- Dividimos el dataset en train y test



## Caso en Big ML de clasificación binaria: *Telecom churn*

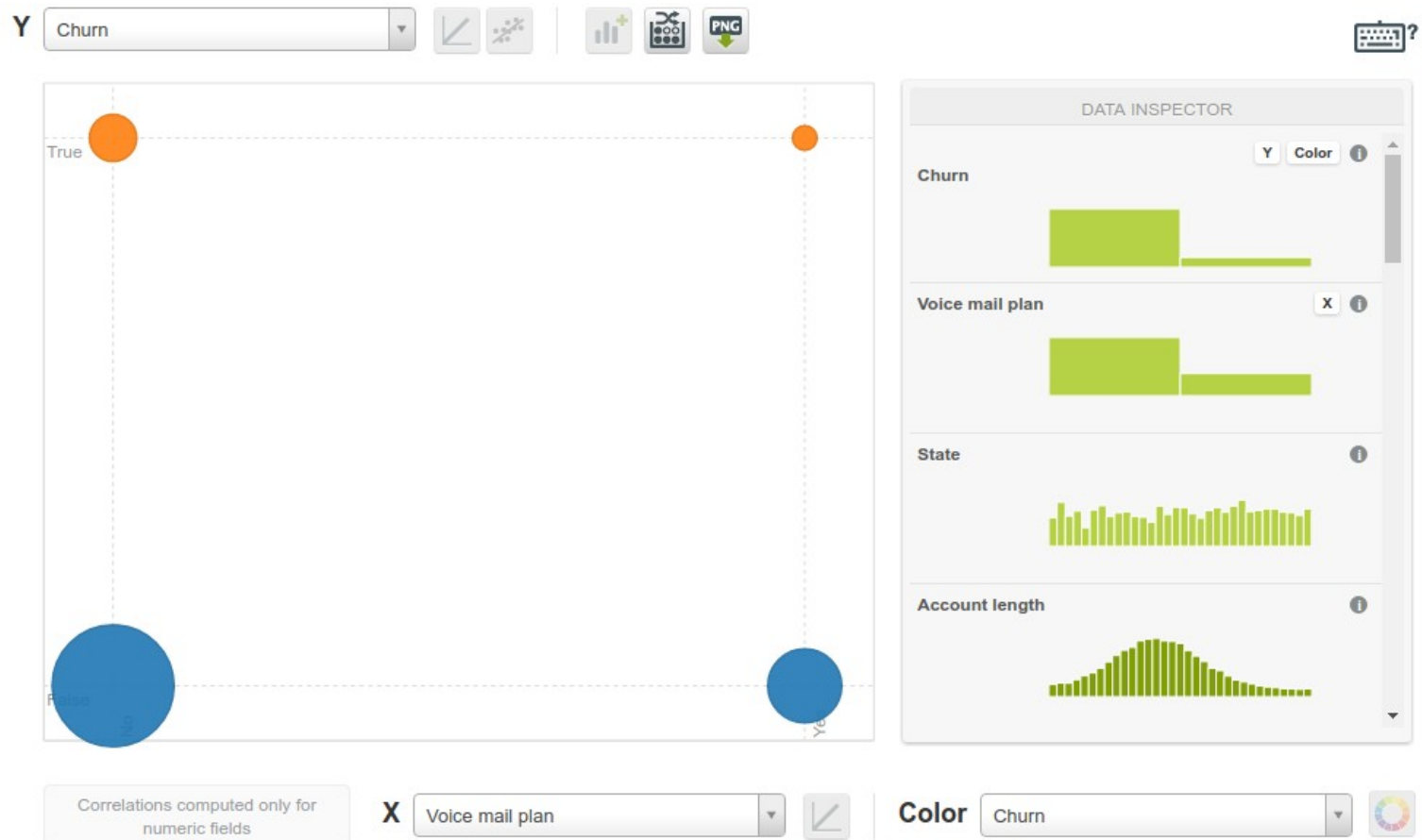
- ¿Qué se puede decir de la relación entre International Plan y Churn?
- ¿Hay diferencias? Comparar con la tabla de contingencia

		International Plan		
		No	Yes	Total
Churn	False	Count 2664 Col% 88.5%	Count 186 Col% 57.6%	Count 2850 Col% 85.5%
	True	Count 346 Col% 11.5%	Count 137 Col% 42.4%	Count 483 Col% 14.5%
	Total	3010	323	3333



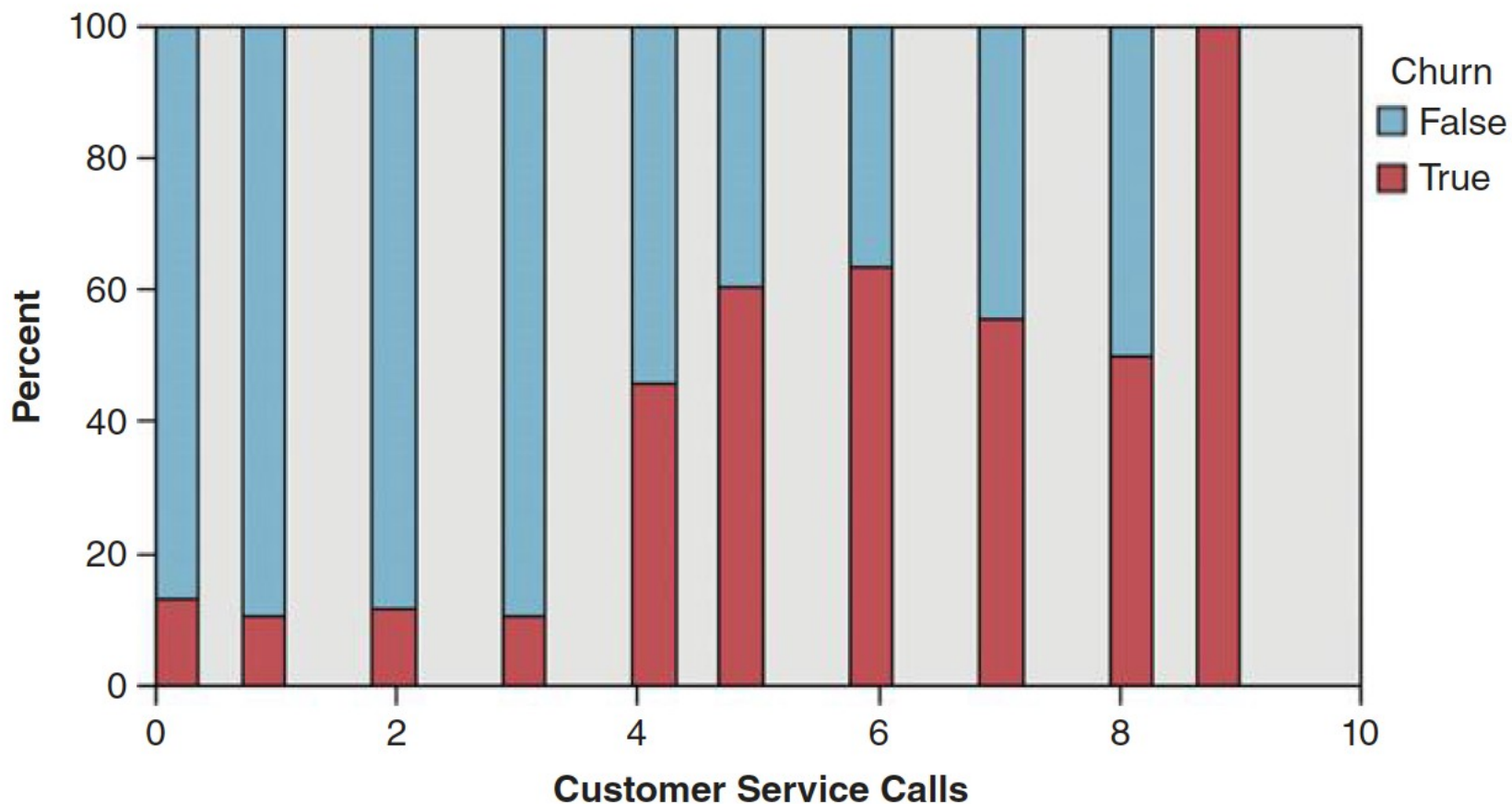
## Caso en Big ML de clasificación binaria: *Telecom churn*

- ¿Qué se podría decir de la relación entre Voice Mail Plan y Churn?
- Parece que la proporción de *churners* es más alta independientemente de VMP.



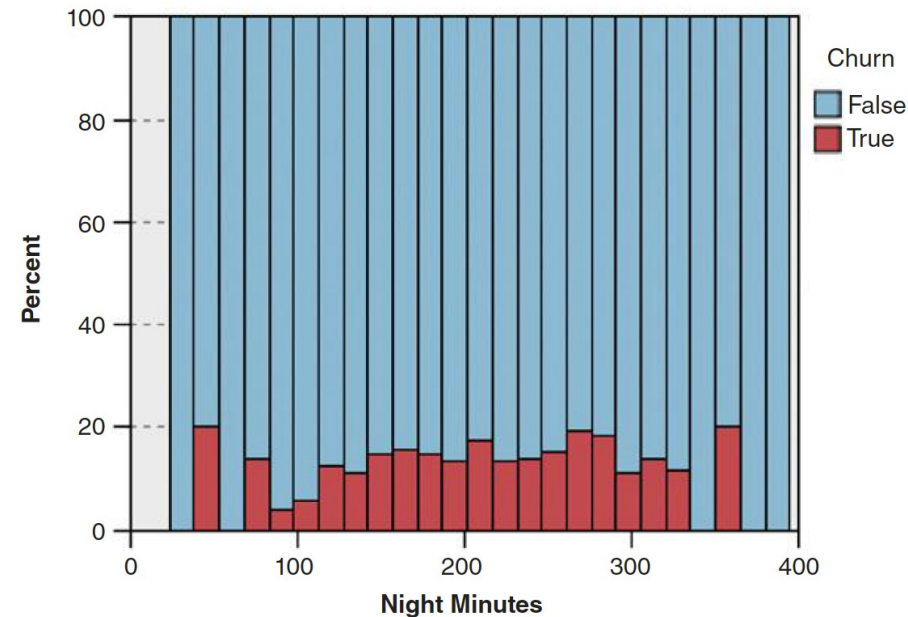
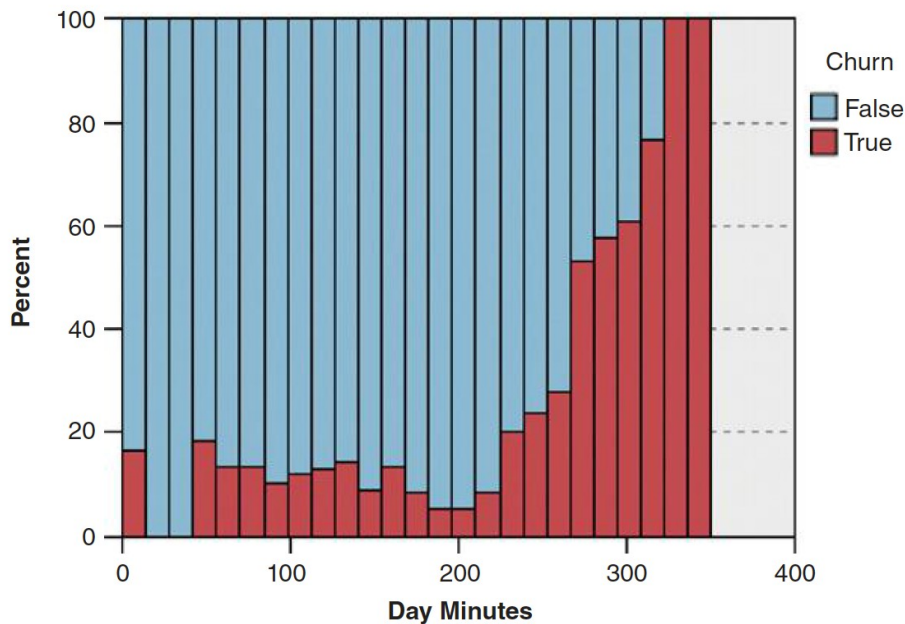
## Caso en Big ML de clasificación binaria: *Telecom churn*

¿Qué se podría decir de la variable **Customer Service Calls**?



## Caso en Big ML de clasificación binaria: *Telecom churn*

Observamos distintas distribuciones de *churn* ó patrones según sea día ó noche



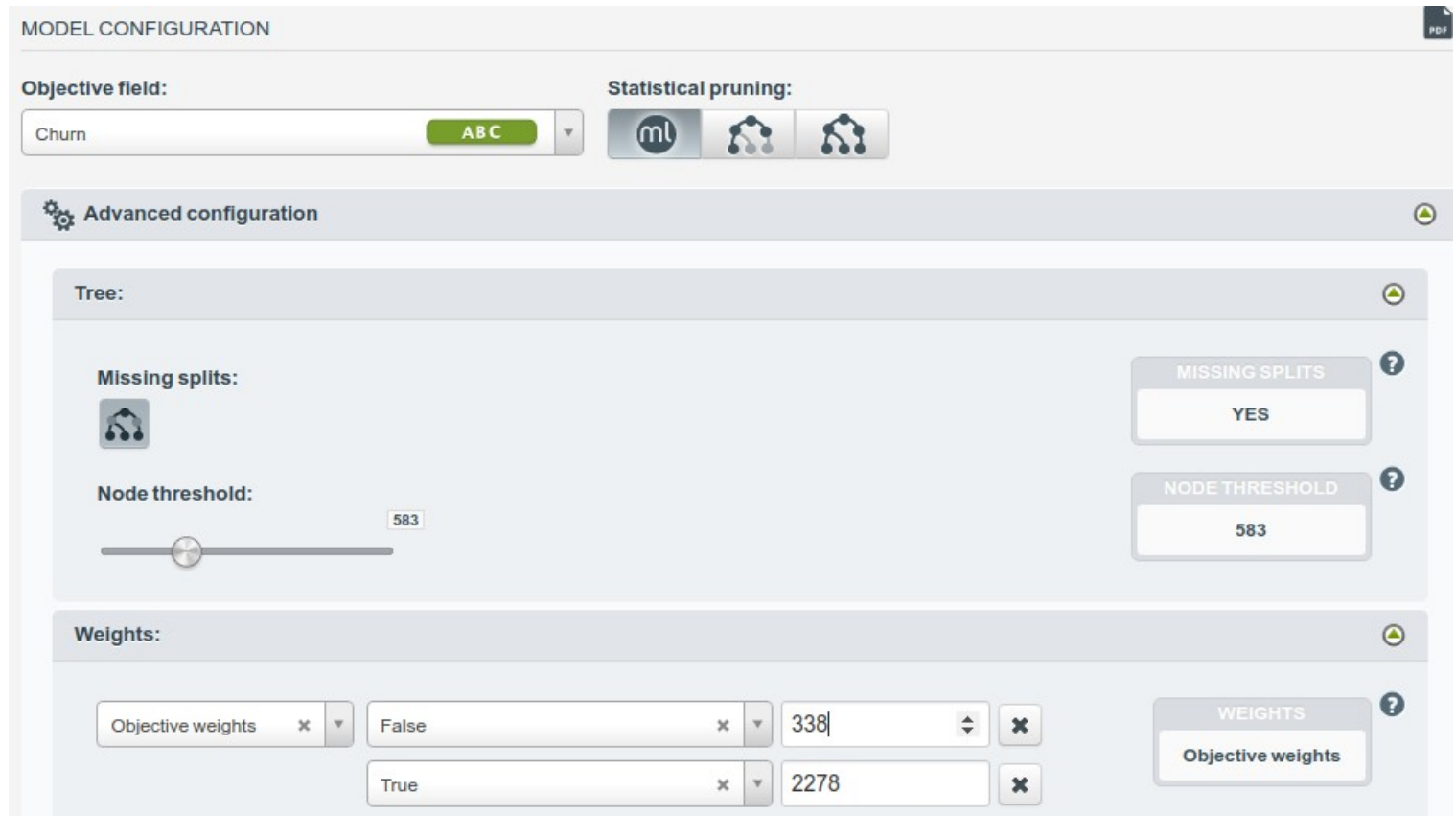
## Caso en Big ML de clasificación binaria: *Telecom churn*

### Preparación de los datos

- En este caso no encontramos valores faltantes, y no parece que haya valores extremos.
- Sin embargo “Area Code” tiene solo tres valores, es extraño.
- **Una observación es que las dos clases están muy desbalanceadas.**
  - **La “paradoja de la exactitud”.** Si los casos de una clase son poco comunes, por ejemplo, un 10%, un clasificador que siempre de la clase mayoritaria tiene un 90% de exactitud ante una muestra.
  - **Podríamos considerar:**
    - *Resampling*: añadir copias sintéticas de ciertos datos.
    - Evaluaciones estratificadas: incluir volúmenes parecidos de las dos clases.
    - No utilizar la *accuracy* sino alguna otra medida: F-score, matrices de confusión, Cohen’s Kappa, curvas ROC.
    - NOTA: En BigML los modelos se equilibran automáticamente con pesos.

## Caso en Big ML de clasificación binaria: *Telecom churn*

Como son clases desequilibradas, introducimos los pesos de la cantidad de la clase contraria al entrenar un modelo en **configure model** .  
En este caso hay 2278 False y 388 True en el conjunto de entrenamiento y ponderamos así:



MODEL CONFIGURATION

Objective field: Churn ABC

Statistical pruning: ml

Advanced configuration

Tree:

Missing splits: YES

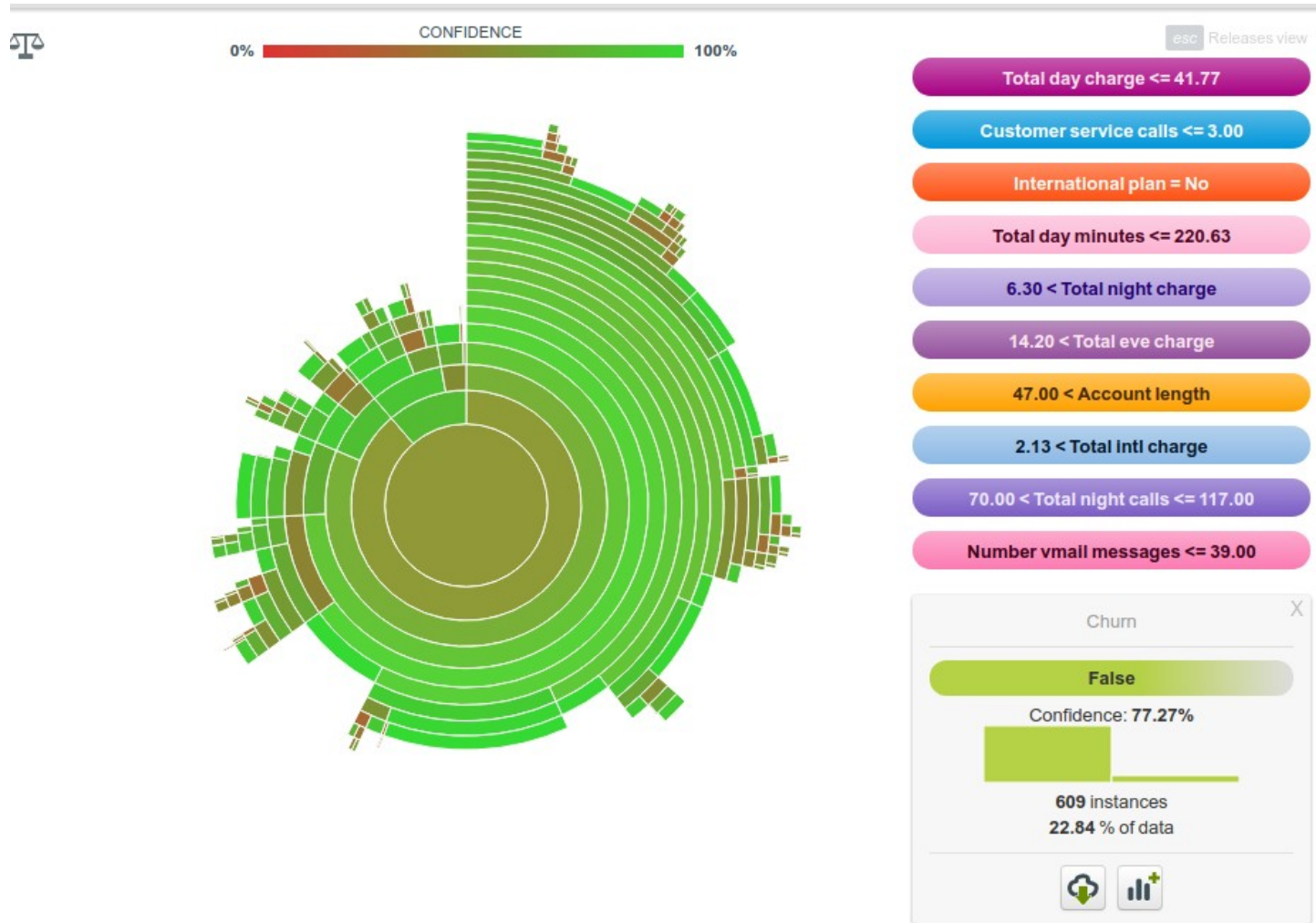
Node threshold: 583

Weights:

Class	Weight
False	338
True	2278

## Caso en Big ML de clasificación binaria: *Telecom churn*

La salida del árbol aporta grados de confianza sobre *churn* en cada hoja



## Caso en Big ML de clasificación binaria: *Telecom churn*

### Evaluamos en el test set

Arriba a la derecha podemos comparar el rendimiento del modelo enfrentándolo a uno aleatorio ó con uno modal.

ACTUAL VS. PREDICTED											
<div><div></div></div>		False	True	ACTUAL		RECALL		F	Phi		
False	529	43	572	92.48%				0.95	0.68		
True	17	78	95	82.10%				0.72	0.68		
PREDICTED	546	121	667	87.29% AVG. RECALL				0.83 AVG. F	0.68 AVG. Phi		
PRECISION	96.89%	64.46%	80.67% AVG. PRECISION		91.00% ACCURACY						

MODEL

91.0%

RANDOM

49.2%

DIFFERENCE

▲41.8%

Accuracy

MODEL

0.7222

RANDOM

0.1909

DIFFERENCE

▲0.5313

F-measure

MODEL

64.5%

RANDOM

12.3%

DIFFERENCE

▲52.1%

Precision

MODEL

82.1%

RANDOM

42.1%

DIFFERENCE

▲40.0%

Recall

MODEL

0.6765

RANDOM

-0.0528

DIFFERENCE

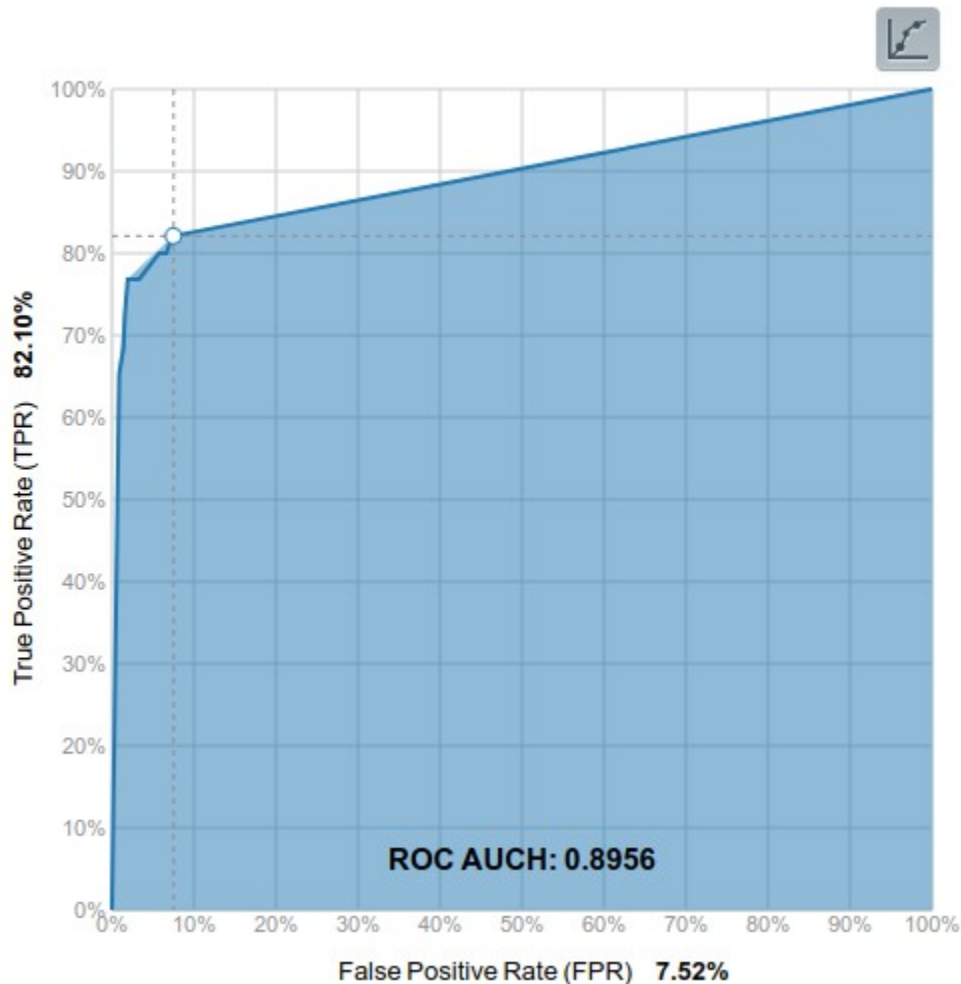
▲0.7292

Phi coefficient



## Caso en Big ML de clasificación binaria: *Telecom churn*

**Evaluamos en el *test set* dando el ROC AUC más métricas**



91.0% Accuracy		0.7222 F-measure	
64.5% Precision	82.1% Recall	0.6765 Phi coefficient	
7.5% FPR	18.1% % positive instances	452.6% Lift	
74.9% K-S statistic	0.6938 Kendall's Tau	0.7149 Spearman's Rho	



## Caso en Big ML de clasificación binaria: *Telecom churn*

### Aplicamos una regresión logística e interpretamos los coeficientes

Imputamos la mediana en los valores faltantes, ponemos pesos en variable objetivo, siempre escalamos los campos y se pueden añadir regularizaciones L1 y L2

The screenshot displays the 'LOGISTIC REGRESSION CONFIGURATION' interface. At the top, the 'Objective field' is set to 'Churn' with a green 'ABC' button. The 'Default numeric value' is set to 'Median'. The 'No missing numerics' field is 'N/A'. The 'Eps' slider is at 0.0001. The 'Stats' button is visible. Below this is the 'Advanced configuration' section. The 'Weights' section shows 'Objective weights' set to 'False' with a value of 338, and 'True' with a value of 2278. The 'Scales' section shows 'Bias' set to  $b_0$  and 'Fields' set to a scale icon. The 'Regularization' section shows 'Regularization' set to 'L2' and 'Strength (c)' set to 1. There are also buttons for 'BIAS' and 'AUTO-SCALED FIELDS' set to 'Yes'.

## Caso en Big ML de clasificación binaria: *Telecom churn*

### Regularización L1 y L2

Para no hacer sobreajuste en la **regresión lineal y logística**, se suelen poner costes a los valores de los coeficientes que se suman a la función de error a minimizar por el algoritmo.

De este modo conseguimos ponderar que los valores que se alejen de 0 de los coeficientes de la regresión logística deben estar justificados con un buen rendimiento explicativo respecto a la variable objetivo.

Hay dos tipos de regularización:

- **L2:** reduce suavemente los coeficientes
- **L1:** hace nulos los coeficientes menos explicativos.
























































Construye lo que se conoce como un modelo *sparse* ó hueco.



## Caso en Big ML de clasificación binaria: *Telecom churn*

### Salida de la regresión logística e interpretación

Tenemos coeficientes y sesgo descritos con valores y significaciones. También un contraste del modelo respecto al modelo constante.

Bias and predictors	Type	False	True
Total intl charge	123 	-13.36090  	13.36090  
Total eve minutes	123 	-4.24046  	4.24046  
Total night minutes	123 	-2.81241  	2.81241  
International plan = Yes	ABC 	-2.42101  	2.42101  
Customer service calls	123 	-0.80461  	0.80461  
Number vmail messages	123 	-0.59937  	0.59937  
Total day minutes	123 	-0.46750  	0.46750  
Total day charge	123 	-0.26325  	0.26325  
Total day calls	123 	-0.11760  	0.11760  
Account length	123 	-0.06654  	0.06654  
Total eve calls	123 	-0.00000  	0.00000  

## Caso en Big ML de clasificación binaria: *Telecom churn*

### Evaluación de la regresión logística

Observamos que el rendimiento de la regresión logística es bastante peor que el del árbol de decisión. ¿ Se podría mejorar este resultado cambiando parámetros? ¿ Qué pasa si se mueve el probability threshold en la pantalla ROC?

ACTUAL VS. PREDICTED				ACTUAL	RECALL	F	Phi
	False	True					
False	434	138	572	75.87%	0.85	0.40	
True	21	74	95	77.90%	0.48	0.40	
PREDICTED	455	212	667	76.88% AVG. RECALL	0.66 AVG. F	0.40 AVG. Phi	
PRECISION	95.38%	34.91%	65.14% AVG. PRECISION	76.16% ACCURACY			

MODEL

76.2%

MODE

85.8%

DIFFERENCE

▼-9.6%

Accuracy

MODEL

0.4821

MODE

0

DIFFERENCE

▲0.4821

F-measure

MODEL

34.9%

MODE

0.0%

DIFFERENCE

▲34.9%

Precision

MODEL

77.9%

MODE

0.0%

DIFFERENCE

▲77.9%

Recall

MODEL

0.4036

MODE

0

DIFFERENCE

▲0.4036






Phi coefficient

## Caso en Big ML de clasificación binaria: *Telecom churn*

### Reejucución y evaluación de la regresión logística

- No imputamos valores faltantes
- No ponemos pesos en la variable objetivo

Ejecutamos de nuevo la regresión logística y reevaluamos obtniendo mejores resultados. Aun así no mejora los del modelo de árbol.

International plan = Yes	ABC	-1.94170	 $\sigma$	1.94170	 $\sigma$
Customer service calls	123 	-0.65633	 $\sigma$	0.65633	 $\sigma$

¿ Cómo cambia la probabilidad de abandono de un cliente con plan internacional frente a un cliente sin el mismo?






Por cada llamada a atención al cliente, ¿ cuánto aumenta la probabilidad de abandono?

## Caso en Big ML de clasificación binaria: *Telecom churn*

### Reejecución y evaluación de la regresión logística

- No imputamos valores faltantes
- No ponemos pesos en la variable objetivo

Ejecutamos de nuevo la regresión logística y reevaluamos obteniendo mejores resultados. Aun así no mejora los del modelo de árbol.

International plan = Yes	ABC	-1.94170	 $\sigma$	1.94170	 $\sigma$
Customer service calls	123 	-0.65633	 $\sigma$	0.65633	 $\sigma$

¿ Cómo cambia la probabilidad de abandono de un cliente con plan internacional frente a un cliente sin el mismo?

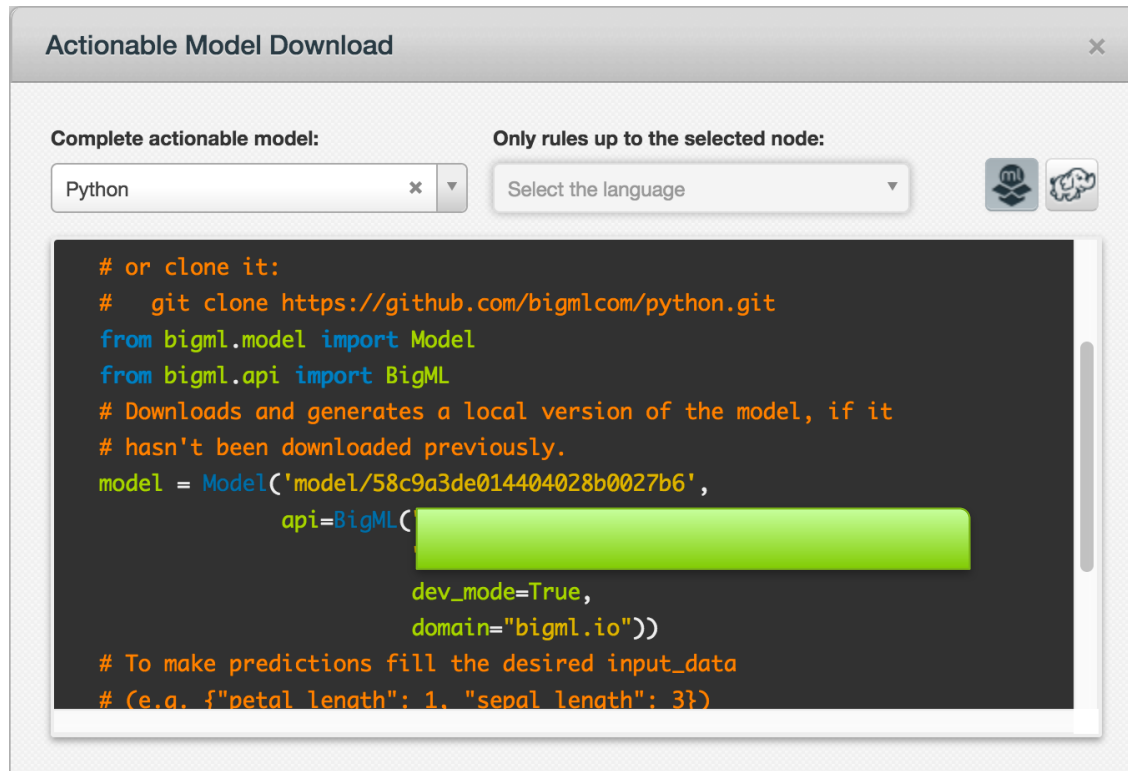
Por cada llamada a atención al cliente, ¿ cuánto aumenta la probabilidad de abandono?

## Caso en Big ML de clasificación binaria: *Telecom churn*

### Despliegue del modelo

Si usamos la API de Big ML para crear un modelo tenemos las siguientes opciones para desplegarlo y usarlo:

- Generar código para integrar en los sistemas actuales.
- Generar código para ejecutar las predicciones en la nube.



The screenshot shows a window titled "Actionable Model Download". It has two tabs: "Complete actionable model:" and "Only rules up to the selected node:". The "Complete actionable model:" tab is selected, and it shows a dropdown menu with "Python" selected. To the right of the dropdown is a button with a close icon (x) and a dropdown arrow. Below the tabs is a text area with the following code:

```
# or clone it:
# git clone https://github.com/bigmlcom/python.git
from bigml.model import Model
from bigml.api import BigML
# Downloads and generates a local version of the model, if it
# hasn't been downloaded previously.
model = Model('model/58c9a3de014404028b0027b6',
              api=BigML(
                  # [Redacted]
                  dev_mode=True,
                  domain="bigml.io"))
# To make predictions fill the desired input_data
# (e.g. {"petal length": 1, "sepal length": 3})
```

## Clasificación múltiple

Se propone como actividad ó ejercicio el análisis completo de otto\_dataset.csv proveniente del famoso concurso de Kaggle con 10 categorías a predecir

<https://www.kaggle.com/c/otto-group-product-classification-challenge>

En Kaggle, los participantes obtienen entorno a un 84% accuracy.

Vemos la entrevista de los enfoques del ganador y el segundo:

- <https://www.kaggle.com/c/otto-group-product-classification-challenge/discussion/14335>
- <http://blog.kaggle.com/2015/06/09/otto-product-classification-winners-interview-2nd-place-alexander-guschin/>

¿ Qué se puede hacer para mejorar los rendimientos una vez llegado a un límite de precisión?

# Ingeniería de atributos

Damos ejemplos en la pizarra de aplicaciones que permiten dividir los datos incorporando estas técnicas que suben de dimensión convenientemente los mismos.



## Técnicas avanzadas de combinación de modelos

Además de la ingeniería de atributos y combinaciones de modelos más básicas, se pueden construir estructuras bastante complejas que suben el rendimiento considerablemente.

Eso sí, siempre con el **principio de parsimonia** en mente...

Una estructura interesante la montó el ganador del [concurso de clasificación de hojas](#)

