

Data centric: cuidando los ingredientes

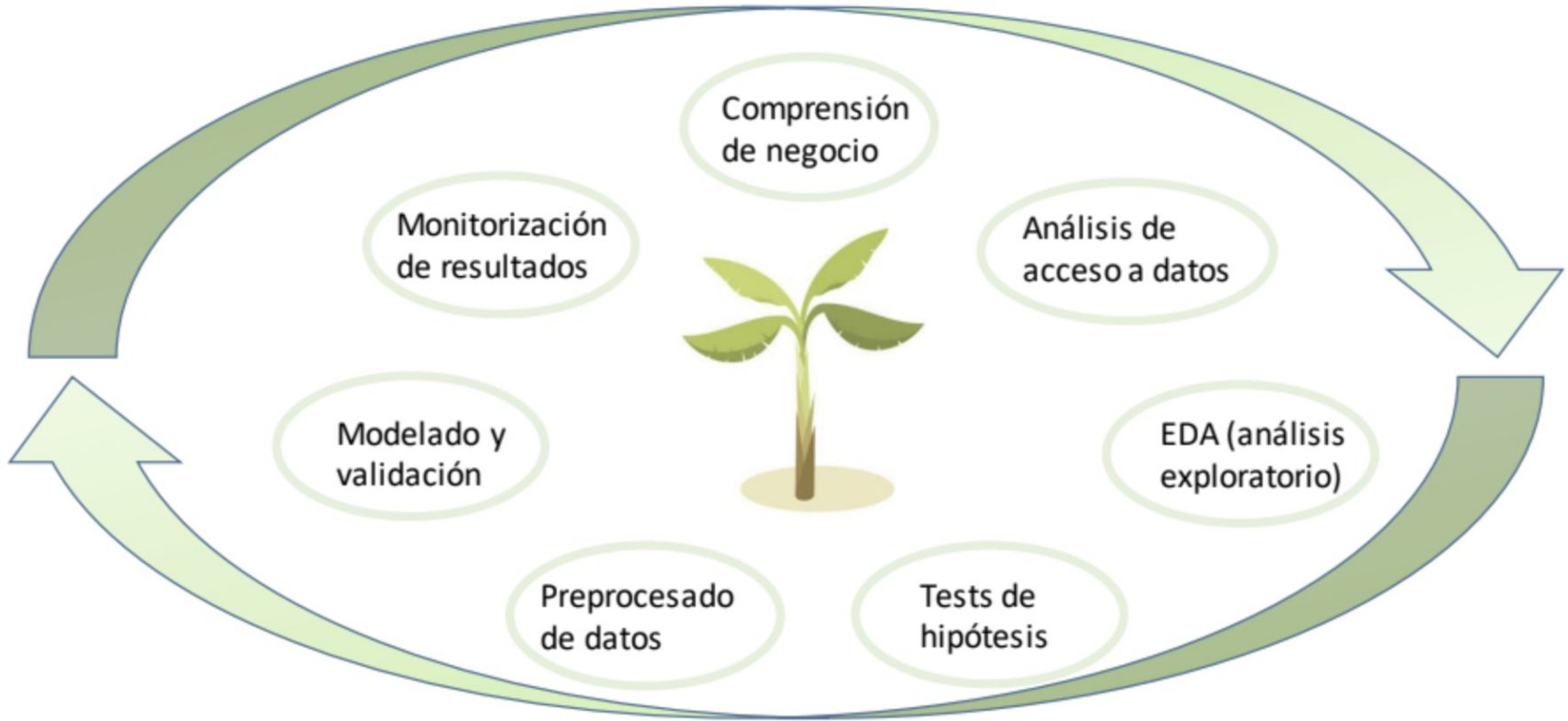
Datos de contacto:

 <https://www.linkedin.com/in/german-cm/>

Data Science <https://datascience.stackexchange.com/users/83791/german-c-m>

 germanthro86@gmail.com

CICLO DE VIDA DE MODELOS en el contexto de ciencia de datos



CALIDAD DEL DATO: lo que imaginas



CALIDAD DEL DATO:

lo que tienes



CALIDAD DEL DATO:

algunas estimaciones



Gartner: se estima que alrededor de un 60% de las empresas no miden el coste anual asociado a un bajo nivel de calidad de sus datos

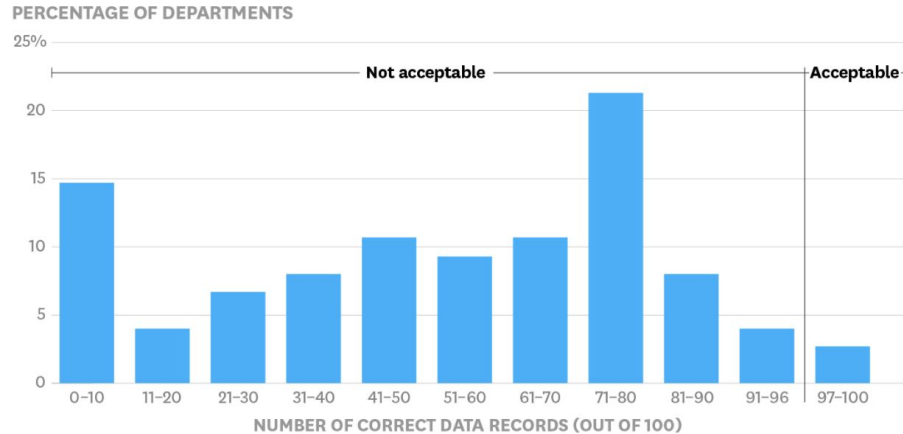
CALIDAD DEL DATO:

algunas estimaciones

👉 Gartner: se estima que alrededor de un 60% de las empresas no miden el coste anual asociado a un bajo nivel de calidad de sus datos

👉 Harvard Business Review: solamente en torno a un 3% de los datos disponibles entre las empresas consultadas en un reciente estudio cumple el umbral mínimo considerado como acceptable:

Analytics And Data Science | Only 3% of Companies' Data Meets Basic Quality Standards



SOURCE TADHG NAGLE ET AL.

© HBR.ORG

CALIDAD DEL DATO:

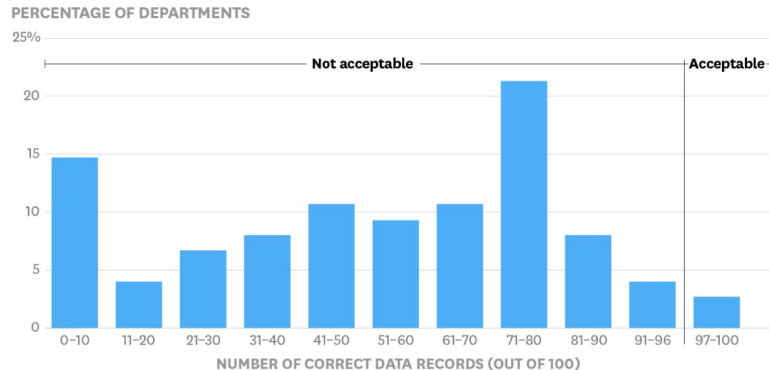
algunas estimaciones

👉 Gartner: se estima que alrededor de un 60% de las empresas no miden el coste anual asociado a un bajo nivel de calidad de sus datos

👉 Harvard Business Review: solamente en torno a un 3% de los datos disponibles entre las empresas consultadas en un reciente estudio cumple el umbral mínimo considerado como aceptable:

👉 Harvard Business Review: sobre la mitad de los departamentos consultados no llega a un 57% con datos considerados correctos

Analytics And Data Science | Only 3% of Companies' Data Meets Basic Quality Standards



CALIDAD DEL DATO:

Qué se entiende por datos correctos

CALIDAD DEL DATO:

Qué se entiende por datos correctos

- completitud: se dispone de suficiente información para definir cada muestra/elemento de nuestra población de estudio

CALIDAD DEL DATO:

Qué se entiende por datos correctos

- completitud: se dispone de suficiente información para definir cada muestra/elemento de nuestra población de estudio
- fiabilidad/veracidad: podemos confiar en que los datos representan el sistema que vamos a analizar/modelar?

CALIDAD DEL DATO:

Qué se entiende por datos correctos

- completitud: se dispone de suficiente información para definir cada muestra/elemento de nuestra población de estudio
- fiabilidad/veracidad: podemos confiar en que los datos representan el sistema que vamos a analizar/modelar?
- consistencia: son los datos uniformes en la organización? O la misma información está en varias fuentes pero con distinto valor? Influye en la veracidad

CALIDAD DEL DATO:

Qué se entiende por datos correctos

- completitud: se dispone de suficiente información para definir cada muestra/elemento de nuestra población de estudio
- fiabilidad/veracidad: podemos confiar en que los datos representan el sistema que vamos a analizar/modelar?
- consistencia: son los datos uniformes en la organización? O la misma información está en varias fuentes pero con distinto valor? Infiuye en la veracidad
- integridad: se actualizan correctamente los cambios y nuevos datos a lo largo de las fuentes de datos de la compañía? O hay fuentes bien mantenidas y otras obsoletas?

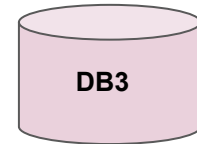
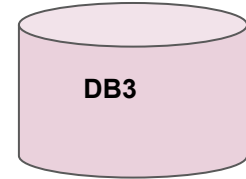
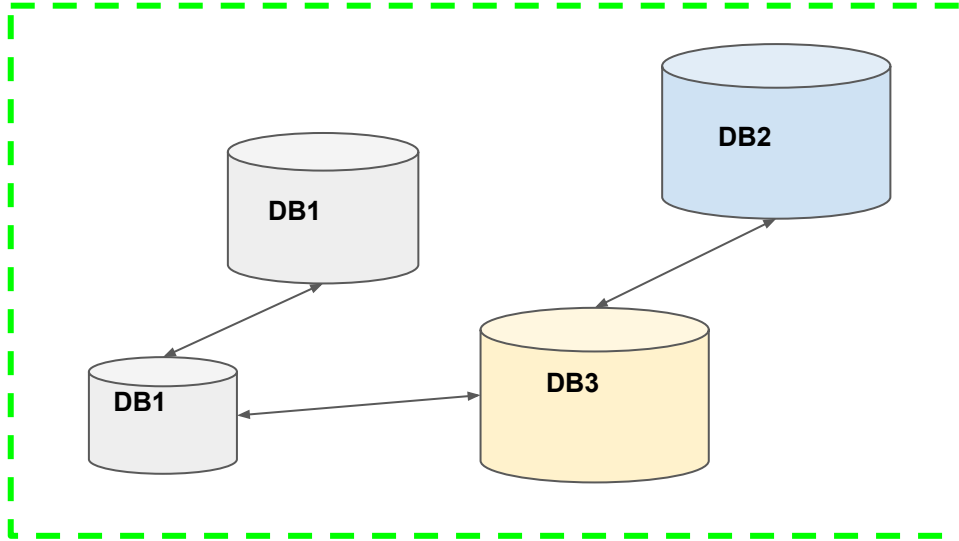
CALIDAD DEL DATO:

Qué se entiende por datos correctos

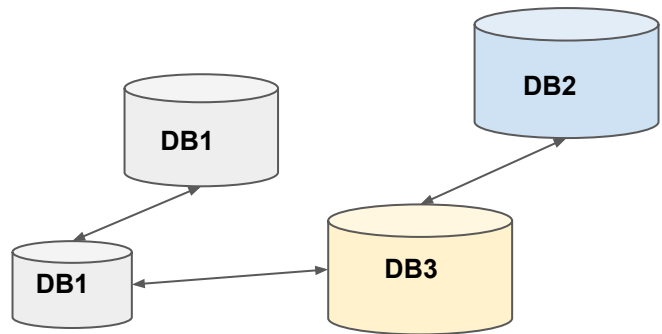
- completitud: se dispone de suficiente información para definir cada muestra/elemento de nuestra población de estudio
- fiabilidad/veracidad: podemos confiar en que los datos representan el sistema que vamos a analizar/modelar?
- consistencia: son los datos uniformes en la organización? O la misma información está en varias fuentes pero con distinto valor? Influye en la veracidad
- integridad: se actualizan correctamente los cambios y nuevos datos a lo largo de las fuentes de datos de la compañía? O hay fuentes bien mantenidas y otras obsoletas?
- linaje de datos: somos capaces de identificar el proceso que generó ciertos datos? Podemos saber cuándo se generaron?

CALIDAD DEL DATO

Identificación de las fuentes de datos necesarias



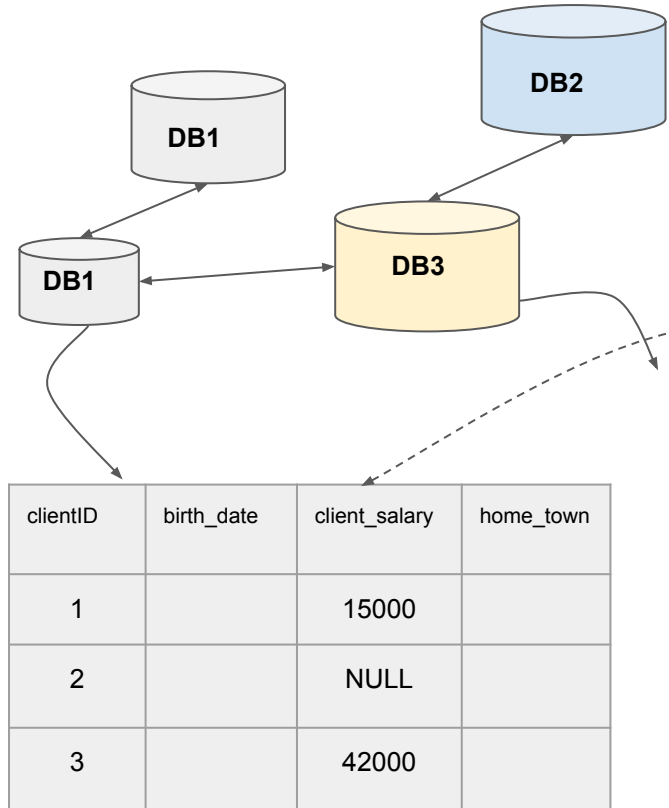
CALIDAD DEL DATO



PROBLEMAS FRECUENTES EJEMPLOS:

- datos distribuidos en silos aislados: posible **mala arquitectura** de datos: reproducir el **esquema de relación de entidades** de las tablas que contienen los datos necesarios para nuestro caso de uso (ausencia de FKs que facilite la comprensión en caso de ser relacional)

CALIDAD DEL DATO

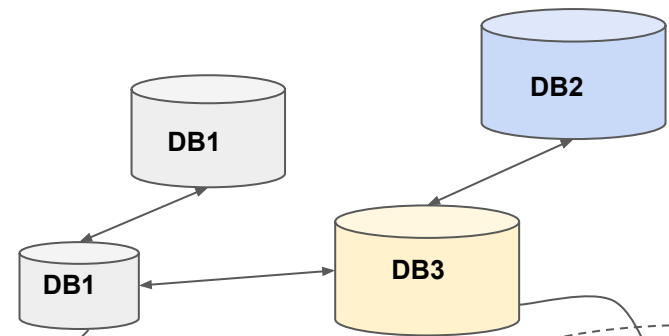


PROBLEMAS FRECUENTES EJEMPLOS:

- datos distribuidos en silos aislados
- **consistencia**: misma info en diferentes tablas y con distintos procesos de carga → posibles problemas de sincronización entre fuentes de datos

clientDOC	clientJob	client_salary
1		25000
2		30000
3		3000

CALIDAD DEL DATO



clientID	birth_date	client_salary	home_town

clientDOC	clientJob	clientAge
		✖

join

clientID	birth_date	client_salary	clientJob	calc_clientAge



PROBLEMAS FRECUENTES

EJEMPLOS:

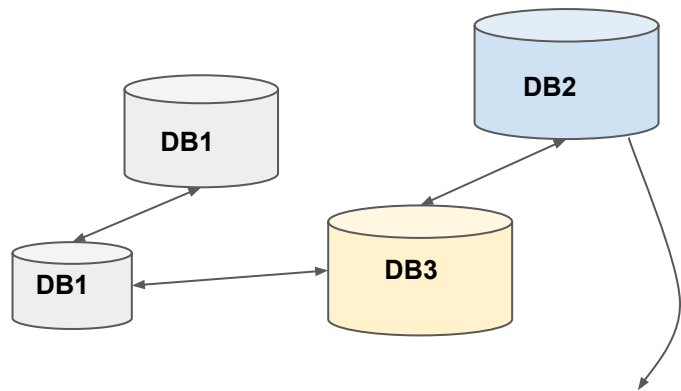
- datos distribuidos en silos aislados
- **consistencia**: misma info en diferentes tablas y con distinta veracidad
- **veracidad** no asegurada: momento de almacenamiento? Origen del dato? Disponemos de

más fuentes

no todos los datos suelen estar actualizados:
CONFIRMAR FUENTE GENERADORA de datos O
RE-CALCULARLOS NOSOTROS

función que
calcula edades

CALIDAD DEL DATO



PROBLEMAS FRECUENTES EJEMPLOS:

- datos distribuidos en silos aislados
- valores duplicados
- **veracidad** no asegurada: momento de almacenamiento? Origen del dato?
- **completitud**: qué datos consideramos el *core* de nuestro caso de uso? De cuáles podemos prescindir y cuáles son obligatorios para el negocio?

ejemplo: info productos/movimientos bancarios
clientes: INFO CORE

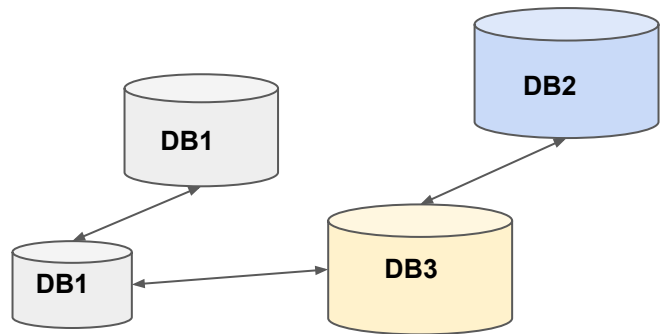
cliente_1							
cliente_2							NULL
cliente_3							

info extra uso de aplicación móvil de clientes

cliente_1			
cliente_3			

no eliminar registros core!

CALIDAD DEL DATO



PROBLEMAS FRECUENTES EJEMPLOS:

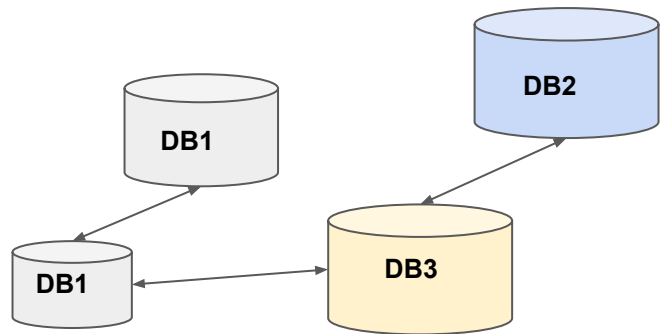
- datos distribuidos en silos aislados
- valores duplicados
- **veracidad** no asegurada: momento de almacenamiento? Origen del dato?
- **completitud**: qué datos consideramos el *core* de nuestro caso de uso? De cuáles podemos prescindir o cuáles se pueden filtrar?
- **frecuencia de muestreo** adecuada? No siempre disponemos de la granularidad deseada

MUESTREO MENSUAL



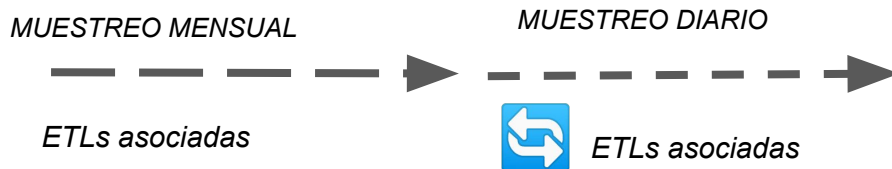
ETLs asociadas

CALIDAD DEL DATO

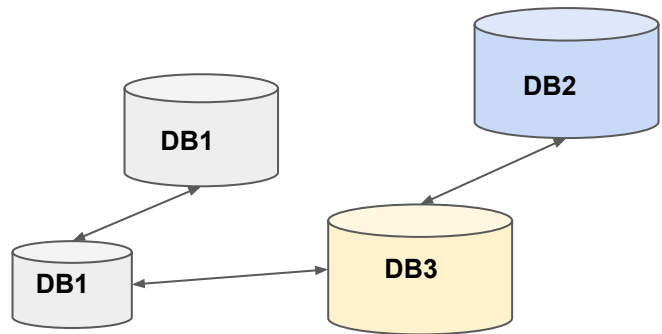


PROBLEMAS FRECUENTES EJEMPLOS:

- datos distribuidos en silos aislados
- valores duplicados
- **veracidad** no asegurada: momento de almacenamiento? Origen del dato?
- **completitud**: qué datos consideramos el *core* de nuestro caso de uso? De cuáles podemos prescindir o cuáles se pueden filtrar?
- **frecuencia de muestreo** adecuada? No siempre disponemos de la granularidad deseada



CALIDAD DEL DATO



PROBLEMAS FRECUENTES EJEMPLOS:

- datos distribuidos en silos aislados
- valores duplicados
- **veracidad** no asegurada: momento de almacenamiento? Origen del dato?
- **completitud**: qué datos consideramos el *core* de nuestro caso de uso? De cuáles podemos prescindir o cuáles se pueden filtrar?
- **frecuencia de muestreo** adecuada? No siempre disponemos de la granularidad deseada

MUESTREO MENSUAL

MUESTREO DIARIO

MUESTREO HORARIO

ETLs asociadas

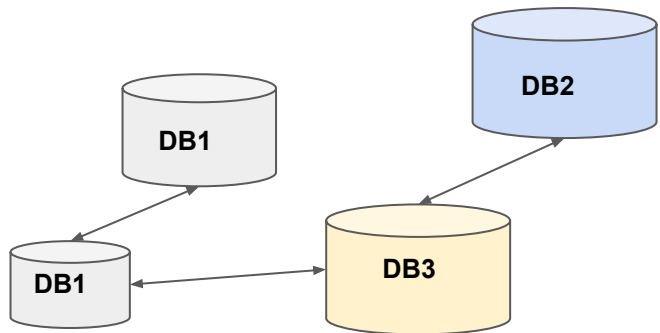


ETLs asociadas



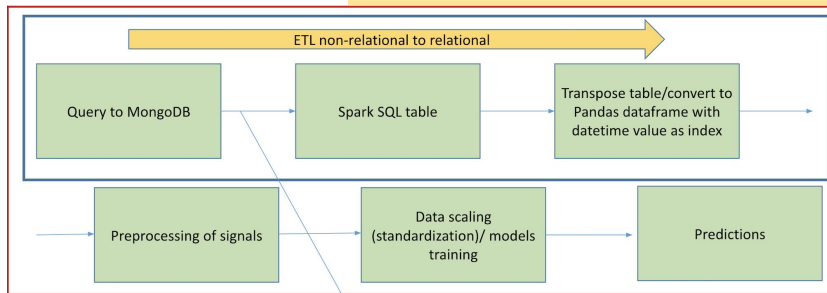
ETLs asociadas

CALIDAD DEL DATO



PROBLEMAS FRECUENTES EJEMPLOS:

- datos distribuidos en silos aislados
- valores duplicados
- **veracidad** no asegurada: momento de almacenamiento? Origen del dato?
- **completitud**: qué datos consideramos el *core* de nuestro caso de uso? De cuáles podemos prescindir o cuáles se pueden filtrar?
- **frecuencia de muestreo** adecuada? No siempre disponemos de la granularidad deseada
- **formato**: conversión a formato deseado: latencias en tiempo real?



Ontology data format

{time, signal_type_1, signal_value_1}	{time, signal_type_2, signal_value_2}	{time, signal_type_3, signal_value_3}
{time, signal_type_4, signal_value_4}	

Desired data format

time	signal_type_1	signal_type_3	signal_type_3	signal_type_4
t_1	value_signal_1	value_signal_2	value_signal_3	value_signal_4

OBJETIVO DE NEGOCIO

OBJETIVO

- solucionar un problema?
 - Minimizar pérdida de clientes
 - Minimizar errores en la producción de piezas de fábrica
 - ...

OBJETIVO

- solucionar un problema?
- crear una nueva funcionalidad?
 - Un sistema que sea capaz de estimar la propensión de un usuario a contratar producto en tiempo real
 - Estimar la propensión de abandonar
 - ...

OBJETIVO

- solucionar un problema?
- crear una nueva funcionalidad?
- mejorar funcionalidad existente?
 - Mejorar el proceso cuando motor de reglas actual no es suficiente
 - Mejorar atención al cliente
 - ...

DEFINICIÓN DE OBJETIVOS

DATOS COMO PRODUCTO



OBJETIVO

DATOS COMO PRODUCTO



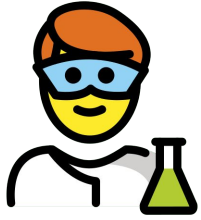
VISIÓN DE PRODUCTO:

- qué funcionalidad que se quiere cubrir
- qué sistema o usuarios utilizarán el resultado: procesos batch VS procesos en tiempo real
- cómo se utilizará el producto resultante: dashboards, modelos, etc
- cómo se evaluará el desempeño del producto resultante
- en qué hitos se puede llevar a cabo (divide y vencerás)
- ...

DEFINICIÓN DE OBJETIVOS

DATOS COMO PRODUCTO

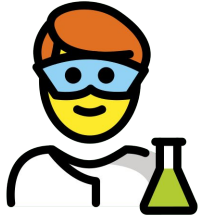




DEFINICIÓN DE OBJETIVOS

DATOS COMO PRODUCTO

¿Se puede abordar con
ciencia de datos?



DEFINICIÓN DE OBJETIVOS

DATOS COMO PRODUCTO

¿Se puede abordar con ciencia de datos?

SÍ

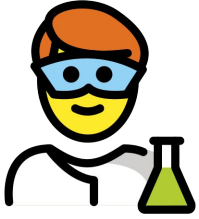
NO

por no disponer de datos a corto plazo?

- qué disponibilidad de datos hay? Datos internos o dependencia de proveedores externos?
- existe documentación de origen y arquitectura de datos?
- qué histórico existe?
- es posible contactar con los equipos encargados de almacenar esos datos?

DEFINICIÓN DE OBJETIVOS

DATOS COMO PRODUCTO



¿Se puede abordar con ciencia de datos?

SÍ

- qué disponibilidad de datos hay? Datos internos o dependencia de proveedores externos?
- existe documentación de origen y arquitectura de datos?
- qué histórico existe?
- es posible contactar con los equipos encargados de almacenar esos datos?

NO

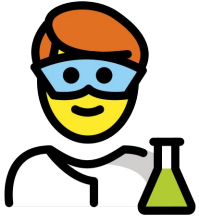
por no disponer de datos a corto plazo?

SÍ

- petición de recogida de datos
- acuerdos con proveedores externos
- simulación de datos: KDE...

DEFINICIÓN DE OBJETIVOS

DATOS COMO PRODUCTO



¿Se puede abordar con ciencia de datos?

SÍ

- qué disponibilidad de datos hay? Datos internos o dependencia de proveedores externos?
- existe documentación de origen y arquitectura de datos?
- qué histórico existe?
- es posible contactar con los equipos encargados de almacenar esos datos?

NO

por no disponer de datos a corto plazo?

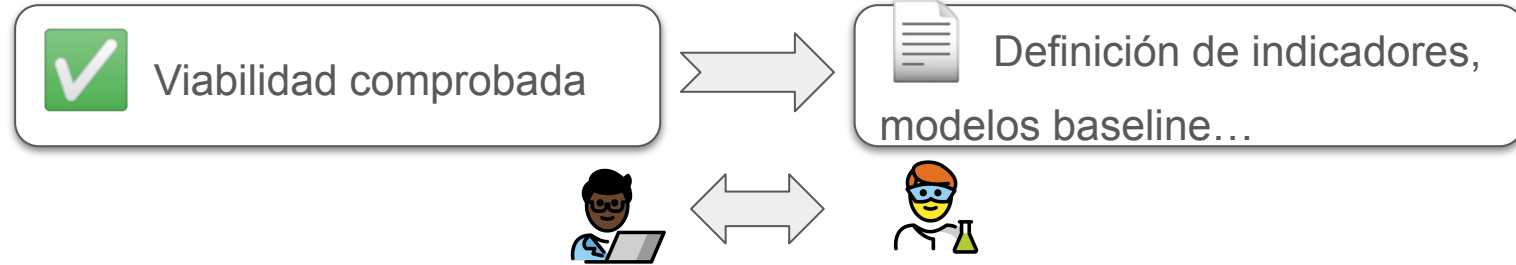
SÍ

- petición de recogida de datos
- acuerdos con proveedores externos
- simulación de datos: KDE...

NO

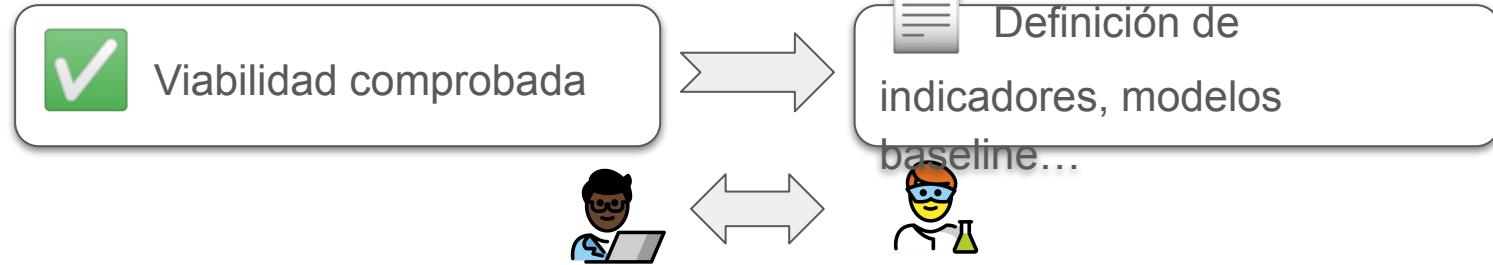
- ⚠ derivar a otros especialistas por no poder enfocarse como proyecto de ciencia de datos

Métricas de negocio *DATOS COMO PRODUCTO*



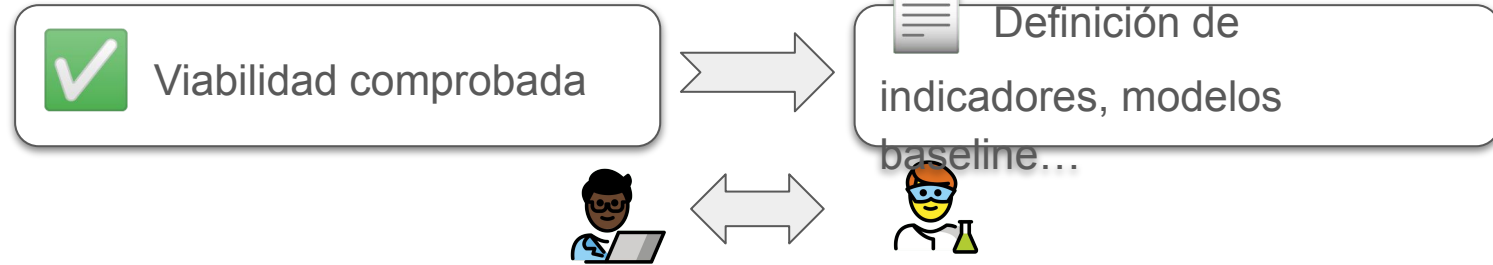
- generación de **KPIs** del caso de uso: *valores de referencia a monitorizar* (nº piezas correctas, nº de nuevos clientes, ahorro mensual de costes...) → diálogo con negocio

KPIs de negocio *DATOS COMO PRODUCTO*



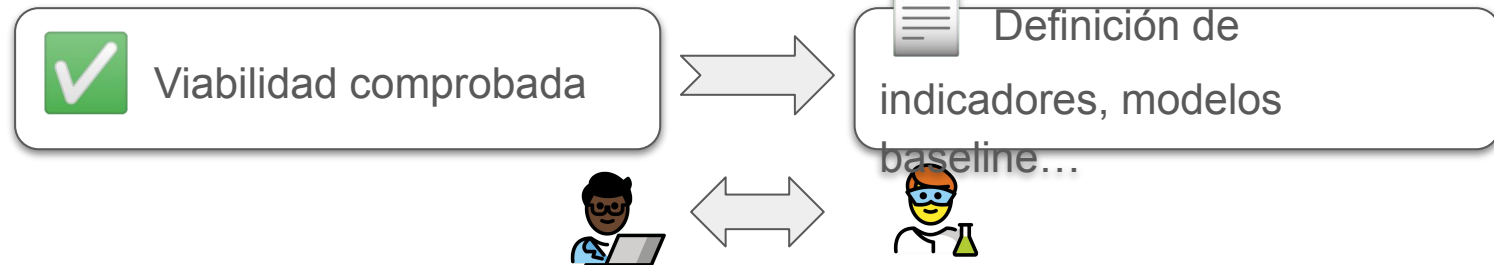
- generación de **KPIs** del caso de uso: *valores de referencia*
- modelos **baseline**: primer acercamiento a los objetivos y referencia para modelos más avanzados

KPIs de negocio *DATOS COMO PRODUCTO*



- generación de **KPIs** del caso de uso: *valores de referencia*
- modelos **baseline**: primer acercamiento a los objetivos y referencia para modelos más avanzados
- motores de **reglas (no subestimar antes de tiempo)**: en base a reglas conocidas de negocio o basadas en análisis exploratorios

KPIs de negocio *DATOS COMO PRODUCTO*

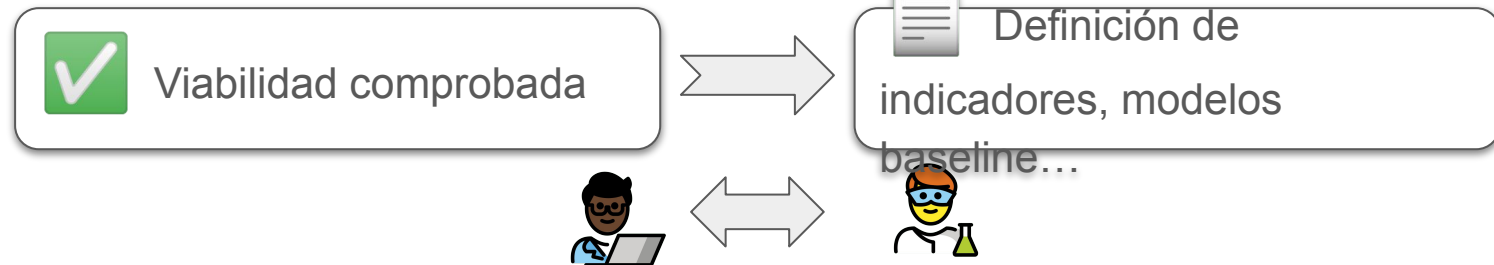




- generación de **KPIs** del caso de uso: *valores de referencia*
- modelos **baseline**: primer acercamiento a los objetivos y referencia para modelos más avanzados
- motores de **reglas (no subestimar antes de tiempo)**: en base a reglas conocidas de negocio o basadas en análisis exploratorios
- **métricas de modelos más avanzados**: modelos de aprendizaje automático
Ejemplo: clasificador → ROC AUC, P-R AUC, F1-score...



KPIs de negocio, reglas conocidas...

DATOS COMO PRODUCTO



- generación de **KPIs** del caso de uso: *valores de referencia*
- modelos **baseline**: primer acercamiento a los objetivos y referencia para modelos más avanzados
- motores de **reglas (no subestimar antes de tiempo)**: en base a reglas conocidas de negocio o basadas en análisis exploratorios
- **métricas de modelos más avanzados**: modelos de aprendizaje automático 
Ejemplo: clasificador → ROC AUC, P-R AUC, F1-score...
- definición de **matriz coste-beneficio**: ¿qué **valor promedio** aporta cada **acierto** de nuestro modelo? ¿Qué **coste promedio** se asocia a cada **error** de dicho modelo? 

valor TN	coste FP
coste FN	valor TP

Errores puntuales, sistemáticos...

- Valores mal informados de **tests de usuarios internos**: suelen darse con poca frecuencia y se pueden filtrar fácilmente
- Errores por **procesos ETL**:
 - son errores sistemáticos que siguen ocurriendo hoy día?
 - se dieron solamente en el pasado pero suponen un porcentaje relevante en nuestro dataset histórico

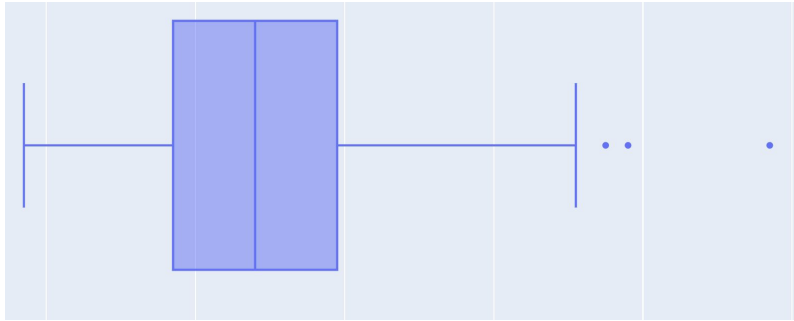
En ambos casos: generar código y tests unitarios en nuestras librerías/utilidades de análisis y modelado

- Datos con **errores que no se pueden solucionar**: Ante la duda, mejor poco y bueno que cantidad con incertidumbre (*Andrew NG: from BIG to GOOD data*):
 - valor erróneo en atributo importante del core dataset: prescindimos del registro completo
 - valor erróneo en atributo de dataset auxiliar: imputamos valor o descartamos atributo si éste presenta alto % de valores erróneos

SEPARANDO EL ORO DE LA ARENA

Detección de valores anómalos:

- exploración con gráficas de caja (box-plots):



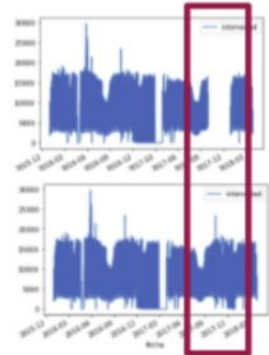
- Z-score (distribuciones gaussianas)
- test de Tukey: valores menores que $Q1 - 1.5IQR$ o mayores que $Q3 + 1.5IQR$

SEPARANDO EL ORO DE LA ARENA



Imputación de **valores ausentes****valores anómalos**:

- valor constante: valor que se quiere forzar en las muestras afectadas
- valor medio/mediano
- valor más frecuente
- interpolación: lineal, polinómica, ...
- sentido común: aprovechar estacionalidad entre otras



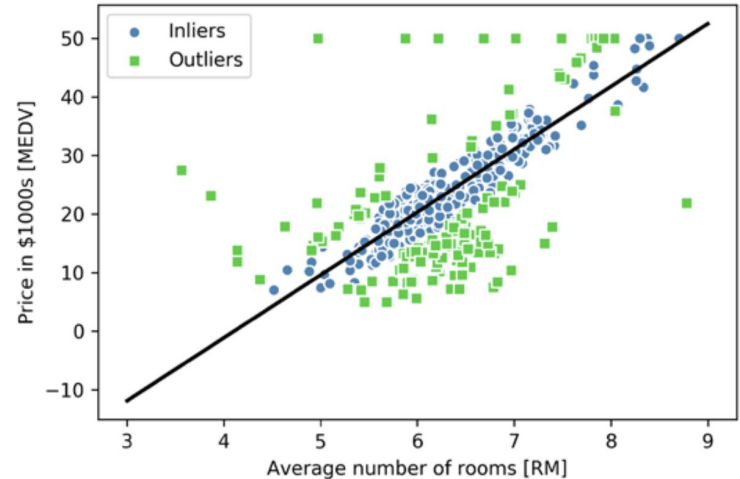
SEPARANDO EL ORO DE LA ARENA

Tratamiento de valores anómalos mediante algoritmos:

- podemos mantenerlos sin considerarlos erróneos
- aplicar algoritmos robustos a dichos valores

Ejemplo: RANSAC (RANDOM SAMPLING CONSENSUS)

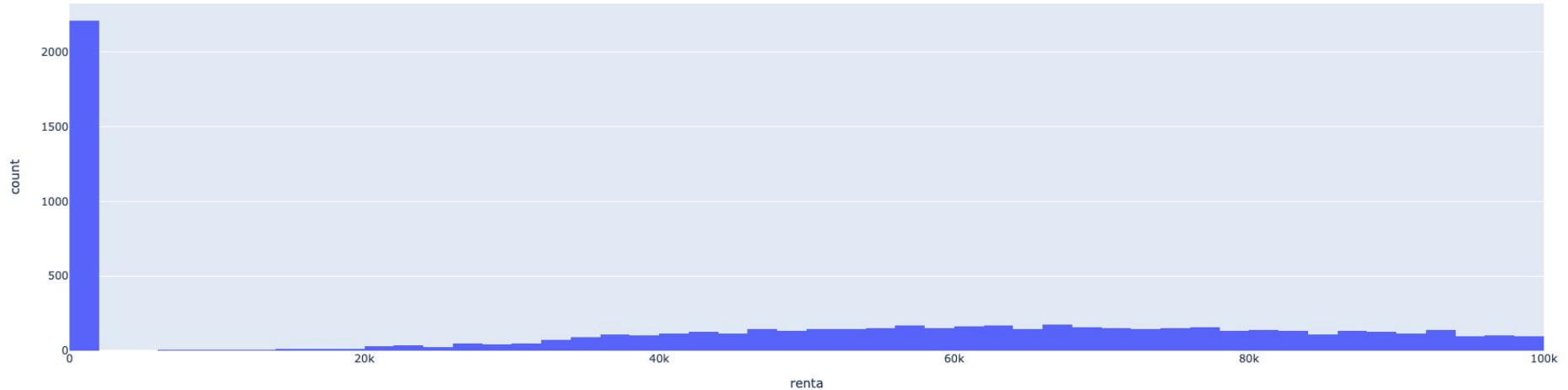
Fundamento: inliers \leq residual_threshold (Median Absolute Deviation por defecto)



SEPARANDO EL ORO DE LA ARENA

Exceso de valores 0:

- tiene sentido para el atributo en cuestión?

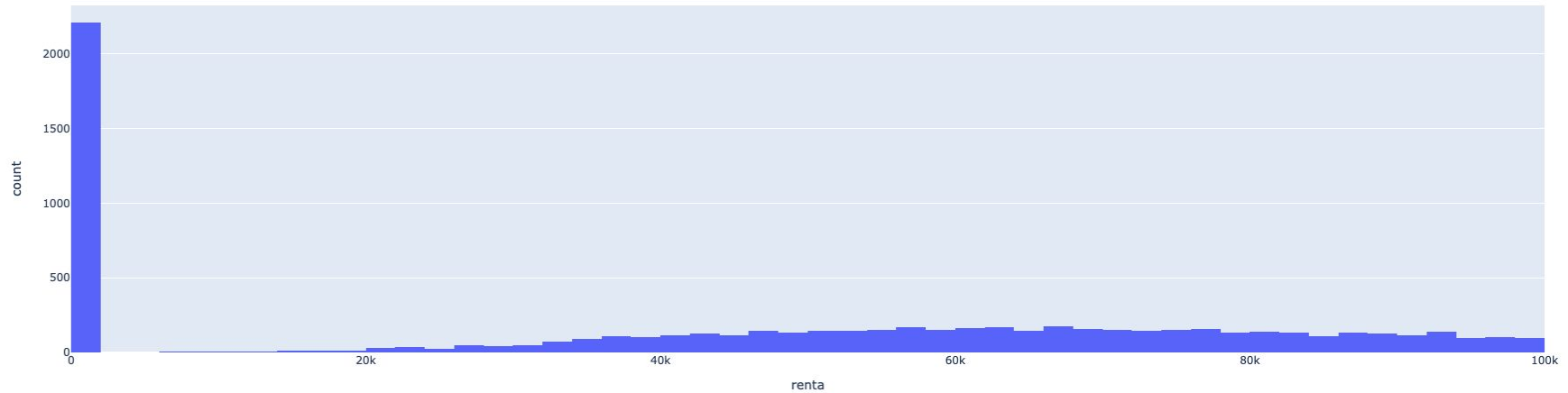


- 💡 son todos estudiantes y podría tener sentido? Tenemos info de su situación laboral? Es un error de valor ausente imputado con 0?

SEPARANDO EL ORO DE LA ARENA

Exceso de valores 0:

- tiene sentido para el atributo en cuestión?



- es este atributo muy relevante para el caso de uso? pertenece al dataset core?
 - sí: utilizar solamente registros con dicho valor informado
 - no: imputar valores como desconocidos y seguir usando dichos registros para modelar

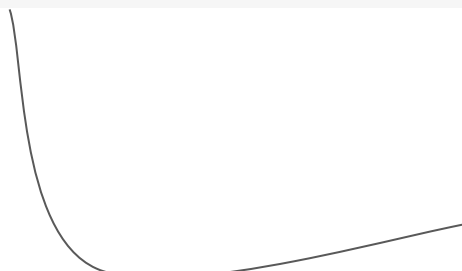
DATA AS A PRODUCT

CONSTRUYENDO PRODUCTO

```
class Dataplotter():  
    def __init__(self, dataframe):  
        self.dataframe_ = dataframe  
  
    def plot_as_histogram(self, column_name, nbins_val=50):  
        try:  
            import plotly.express as px  
  
            fig = px.histogram(self.dataframe_, x=column_name, nbins=nbins_val)  
            fig.show()  
  
        except Exception as exc:  
            print(exc)  
            return exc
```

ganamos en:

- reutilización
- orden
- control de posibles errores
- ...



```
cli_soc_eco_plotter = Dataplotter(clients_socio_eco_df)  
cli_soc_eco_plotter.plot_as_histogram('renta')
```

SEPARANDO EL ORO DE LA ARENA

Sólo los outliers son anómalos?

- **inliers**: valores que caen dentro de cierto rango definido
- **exceso** de **inliers** pueden indicar también valores erróneos

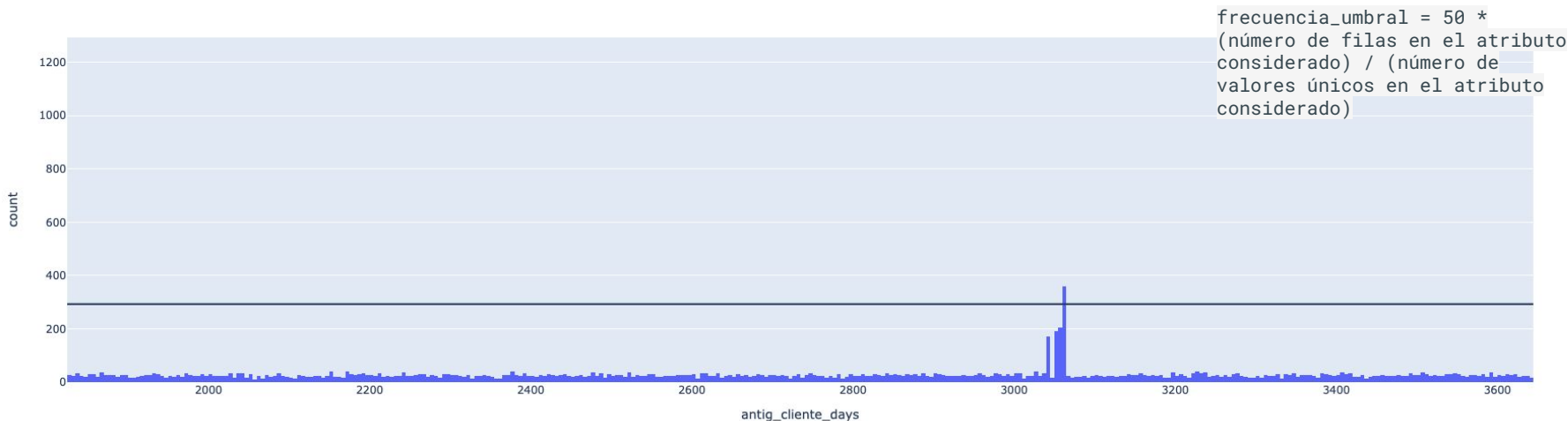


- los mismos valores pueden hacer saltar la alarma o no según la frecuencia de muestreo considerada

SEPARANDO EL ORO DE LA ARENA

Sólo los outliers son anómalos?

- **inliers**: valores que caen dentro de cierto rango definido
- **inliers problemáticos**: inliers anormalmente frecuentes en base a un umbral definido
- **exceso** de **inliers** pueden indicar también valores erróneos
- los mismos valores pueden hacer saltar la alarma o no según la frecuencia de muestreo considerada

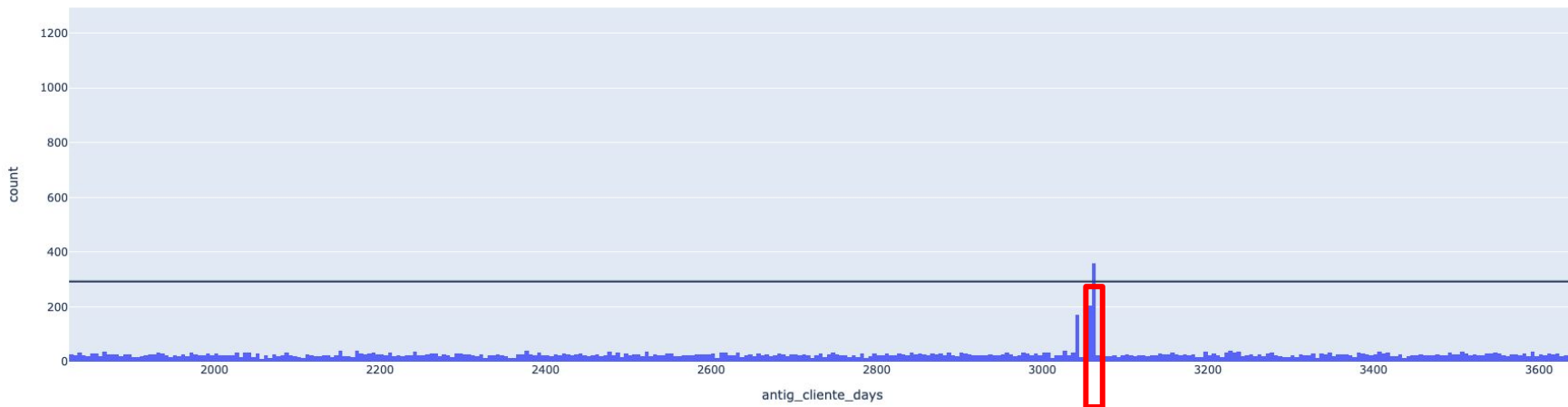


SEPARANDO EL ORO DE LA ARENA

Sólo los outliers son anómalos?

! resultaron coincidir con fechas donde se realizó una migración masiva de datos → entre otros campos de fecha, fecha_alta se actualizó al día de ejecución

! no detectarlos puede llevarnos a ideas de negocio erróneas:
→ ¿qué se hizo bien en esa época de tanta captación?



DATA AS A PRODUCT

CONSTRUYENDO PRODUCTO

```
class Dataplotter():
    def __init__(self, dataframe):
        self.dataframe_ = dataframe

    def plot_as_histogram(self, column_name, nbins_val=50):
        try:
            import plotly.express as px

            fig = px.histogram(self.dataframe_, x=column_name, nbins=nbins_val)
            fig.show()

        except Exception as exc:
            print(exc)
            return exc

    def find_inliers_threshold(self, desired_column):
        try:
            thres_value = int(50 * (len(self.dataframe_[-pd.isna(self.dataframe_[desired_column])
            ])) / len(self.dataframe_[desired_column].unique()))

            return thres_value

        except Exception as exc:
            print(exc)
            return exc

    def plot_hist_values_with_inliers_threshold(self, x_column_name="antig_cliente",
        range_from=0, range_to=1500):
        try:
            import plotly.express as px

            inliers_thres = find_inliers_threshold(self.dataframe_, x_column_name)
            fig = px.histogram(self.dataframe_, x=x_column_name, nbins=int(len(self.dataframe_[x_column_name].unique())/3))
            fig.add_hline(y=inliers_thres)
            fig.update_yaxes(range=[range_from, range_to])
            fig.show()

        except Exception as exc:
            print(exc)
            return exc

cli_soc_eco_plotter = Dataplotter(clients_bank_df)
cli_soc_eco_plotter.plot_hist_values_with_inliers_threshold('antig_cliente_days', range_from=0, range_to=1500)
```



SEPARANDO EL ORO DE LA ARENA

Conteo de accesos de clientes a canal electrónico:

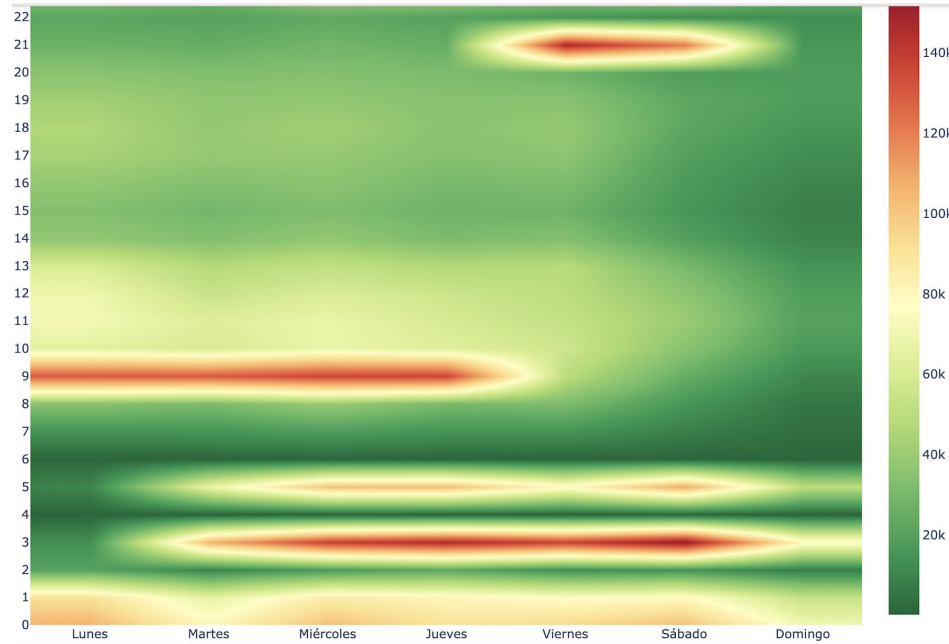
- datos de accesos individuales a nivel horario
- objetivo: obtener número de accesos por cliente para conocer su actividad en canales electrónicos

FECHA_ACCESO	CLIENTE_ID	MES	DÍA	DÍA_SEMANA	HORA
2022-01-18	34	1	18	MARTES	13:00
2022-01-15	5667	1	15	SÁBADO	09:36
2022-01-20	453	1	20	JUEVES	03:00
2022-01-25	34	1	25	MARTES	03:00
...

SEPARANDO EL ORO DE LA ARENA

Comprobación por visualización de accesos agregados por hora y día de la semana:

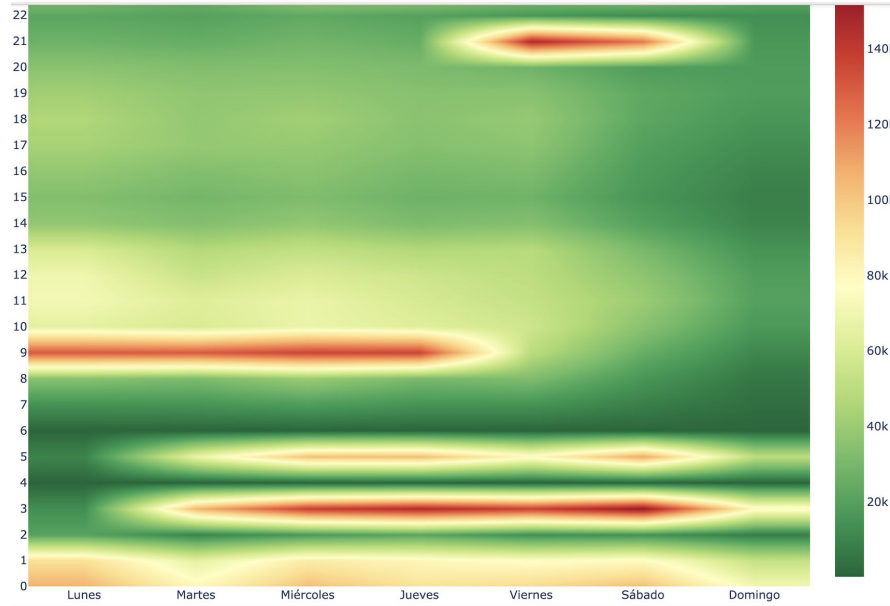
- algo llamativo en este mapa de calor de accesos?
- algo erróneo o puede haber alguna razón?



SEPARANDO EL ORO DE LA ARENA

Consecuencias:

- no conocer las horas reales de conectividad
- contabilizar más accesos de los reales
- peligro de automatizar envío de notificaciones a horas incorrectas



- solución 💡 → filtrado de conexiones que sigan un patrón muy definido (mismas horas en mismos días para muchos usuarios) o por IPs de conexión

VALIDANDO CALIDAD DE NUESTROS DATOS

Software testing + Statistics testing + KPIs checks

- hace esta función de fechas lo que necesito?
- ¿sigue mi atributo la distribución esperada o al menos tiene los valores frecuentes esperados?
- respeta los KPIs conocidos y es coherente con otros valores conocidos? En caso de no ser así, ¿hemos truncado más de lo deseado o hay una razón de peso?

DATOS PROCESADOS: y ahora qué

FEATURE STORE → facilitando la vida del científico de datos

→ almacenamiento de datos analizados, procesados y nuevos atributos generados

- reporting: evolución de número de nuevos productos contratados al mes
- modelado: atributos de entrada ya procesados + nuevos creados
- etiquetado: desacoplado de los atributos de entrada y evolución
- **linaje de datos** → scripts/notebooks generadores de nuestros atributos almacenados + usuarios que generaron dichos atributos + fecha de creación y almacenamiento → **reproducibilidad de experimentos ML**

DATOS PROCESADOS: y ahora qué

FEATURE STORE → facilitando la vida del científico de datos

→ manteniendo nuestros datos analizados, procesados y nuevos atributos generados

- data monitoring → los criterios de negocio pueden cambiar o los resultados de tests de procesos de creación de atributos ser desfavorables → reprocesar

→ **utilidad**: pandas compare para ver qué muestras han podido ser afectadas por cambios de criterio (por ejemplo reprocesado de datos bancarios de clientes)

	ncodpers		MOVIMIENTOS_TARJETA		SALDO_EN_CUENTA	
	self	other	self	other	self	other
0	NaN	NaN	0.0	6.0	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	63.0	457.0	21.0	500.0
3	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN

DATA DRIFT: monitorizando las distribuciones de nuestros datos

PSI: Population Stability Index → métrica para detectar cambios (en base a unos valores umbrales) en la distribución de nuestro atributo de interés a lo largo del tiempo

```
data_drift_tracker_df['PSI'] = data_drift_tracker_df['A-B'] * data_drift_tracker_df['ln(A/B)']  
data_drift_tracker_df
```

	Scoring sample range	Training Percentage (B)	Scoring Percentage (A)	A-B	ln(A/B)	PSI
0	<22	4.4	8.8	4.4	0.693147	3.049848
1	22-24	4.1	10.4	6.3	0.930819	5.864159
2	24-26	3.8	9.6	5.8	0.926762	5.375220
3	26-31	6.9	10.0	3.1	0.371064	1.150297
4	31-38	12.9	9.8	-3.1	-0.274845	0.852019
5	38-43	13.2	10.0	-3.2	-0.277632	0.888422
6	43-48	14.8	11.2	-3.6	-0.278713	1.003368
7	48-54	13.4	10.2	-3.2	-0.272867	0.873174
8	54-64	12.7	9.6	-3.1	-0.279839	0.867501
9	64-116	13.8	10.4	-3.4	-0.282863	0.961733
10	>116	0.0	0.0	0.0	NaN	NaN

```
np.sum(data_drift_tracker_df['PSI'])/100
```

```
0.20885740915133053
```


DATA DRIFT: monitorizando las distribuciones de nuestros datos

PSI: Population Stability Index → métrica para detectar cambios (en base a unos valores umbrales) en la distribución de nuestro atributo de interés a lo largo del tiempo

```
data_drift_tracker_df['PSI'] = data_drift_tracker_df['A-B'] * data_drift_tracker_df['ln(A/B)']
data_drift_tracker_df
```

	Scoring sample range	Training Percentage (B)	Scoring Percentage (A)	A-B	ln(A/B)	PSI
0	<22	4.4	8.8	4.4	0.693147	3.049848
1	22-24	4.1	10.4	6.3	0.930819	5.864159
2	24-26	3.8	9.6	5.8	0.926762	5.375220
3	26-31	6.9	10.0	3.1	0.371064	1.150297
4	31-38	12.9	9.8	-3.1	-0.274845	0.852019
5	38-43	13.2	10.0	-3.2	-0.277632	0.888422
6	43-48	14.8	11.2	-3.6	-0.278713	1.003368
7	48-54	13.4	10.2	-3.2	-0.272867	0.873174
8	54-64	12.7	9.6	-3.1	-0.279839	0.867501
9	64-116	13.8	10.4	-3.4	-0.282863	0.961733
10	>116	0.0	0.0	0.0	NaN	NaN

```
np.sum(data_drift_tracker_df['PSI'])/100
```

```
0.20885740915133053
```

KS de 2 muestras: Kolmogorov-Smirnov test:
bajo la hipótesis nula donde ambas distribuciones proceden de la misma población

```
from scipy import stats
```

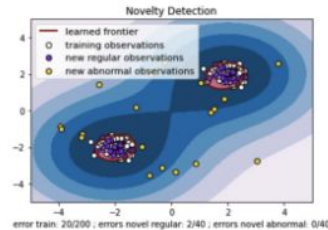
```
stats.ks_2samp(clients_sant_df.age.values, first_clients_sant_df.age.values)
```

```
Ks_2sampResult(statistic=0.1353343516230678, pvalue=0.0)
```

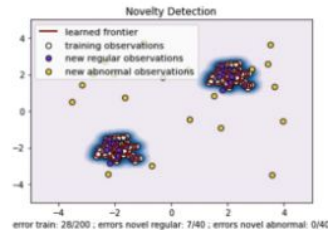
DATA DRIFT: monitorizando nuestros datos

Modelo OC-SVM para monitorizar data drift multivariable:

- permite evaluar la “normalidad” de nuevas muestras respecto a las distribuciones conocidas
- puede utilizarse para alertar de posible drift a nivel multivariable (no solo atributo a atributo)



VS with gamma = 10



where, with this last option, you are overfitting so much.

- conjunto de datos inicial tras procesar: supone nuestro set de entrenamiento, del cual definimos:
→ parámetro ν : fracción de puntos considerados “noveles”
- al cabo de un tiempo, nuevas muestras entrantes en el sistema son evaluadas por este detector de posibles puntos anómalos
- muestras que aporten un alto score como muestra “novel” son sujetos a estudio: ¿hay errores en sus valores?
- si se registran muchas anomalías de este tipo podemos estar ante un cambio significativo (data drift) de nuestro conjunto de datos

Share Edit Delete Flag

answered Sep 27, 2020 at 17:09



German C M

2,264 • 2 • 17

ETIQUETADO DE DATASETS: *directos al objetivo*

No menos importante que cocinar bien la entrada de nuestros modelos es el **buen etiquetado** de nuestros registros según el objetivo perseguido:

- definición **sin ambigüedad**, ejemplo: queremos considerar abandono de un cliente estrictamente cancelación de una cuenta? o queremos añadir también como abandonos clientes que ya no aportan valor? Y qué entendemos por valor de cliente: operar con ciertos productos, mantener un mínimo de saldos, etc
- criterio traducible **en términos de negocio**: se puede explicar con claridad y se puede traducir en números
- qué **anticipación** necesitamos **entre** el momento de aplicación de nuestro modelo (**predicción**) y el momento correspondiente a la **fecha** del valor **predicho** para tener capacidad de acción (ejemplos: campañas, sistema de alertas de una fábrica...)
- cuidado con la fuga de información (**data leakage**): no usar en entrenamiento datos que no estarán disponibles en inferencia
- aplicar técnicas de **data drift** también a la **variable objetivo**: detectar cambios en la distribución del grupo objetivo de nuestro modelo, que puede implicar tener que reentrenar modelos asociados ya sea por:
 - cambio en definición de etiqueta
 - cambio significativo en el % entre positivos y negativos

CALIDAD DE ETIQUETADO

- Etiquetar es un **proceso costoso**
- En gran parte de las ocasiones es manual y hay empresas dedicadas a ello
- A veces nos viene dado por el proceso de negocio

Ejemplo, llamadas a asistentes por voz

→ objetivo: identificar el tema requerido por el usuario sin redirigir a agente humano

→ cómo: mediante técnicas y modelado NLP

- Previamente se tienen definidas unas posibles categorías de asuntos en llamadas
- El usuario llama e indica lo que quiere al asistente automático
- En muchas ocasiones se redirige a un agente
- Cuando el agente atiende una llamada, asigna finalmente una categoría según la conversación mantenida con el cliente

```
| import pandas as pd

calls_dict = {'id': [200, 344, 567, 600, 1042],
              'mensaje_usuario': ['me gustaría abrir una nueva cuenta', 'estado de reclamación',
                                  'quiero realizar un bizum', 'quiero abrir una nueva cuenta',
                                  'quiero una cuenta'],
              'asunto': ['CONSULTA CONDICIONES', 'CANCELACIÓN', 'OPERACIÓN BIZUM', 'CAMBIO TITULARIDAD',
                        'APERTURA DE CUENTA']}

pd.DataFrame(calls_dict)
```

	id	mensaje_usuario	asunto
0	200	me gustaría abrir una nueva cuenta	CONSULTA CONDICIONES
1	344	estado de reclamación	CANCELACIÓN
2	567	quiero realizar un bizum	OPERACIÓN BIZUM
3	600	quiero abrir una nueva cuenta	CAMBIO TITULARIDAD
4	1042	quiero una cuenta	APERTURA DE CUENTA

- → analizar n-gramas (grupos de 1, 2, 3... palabras) más frecuentes por asunto e intentar identificar posibles malos etiquetados
- → validar criterio de etiquetado con los encargados de asignarlas

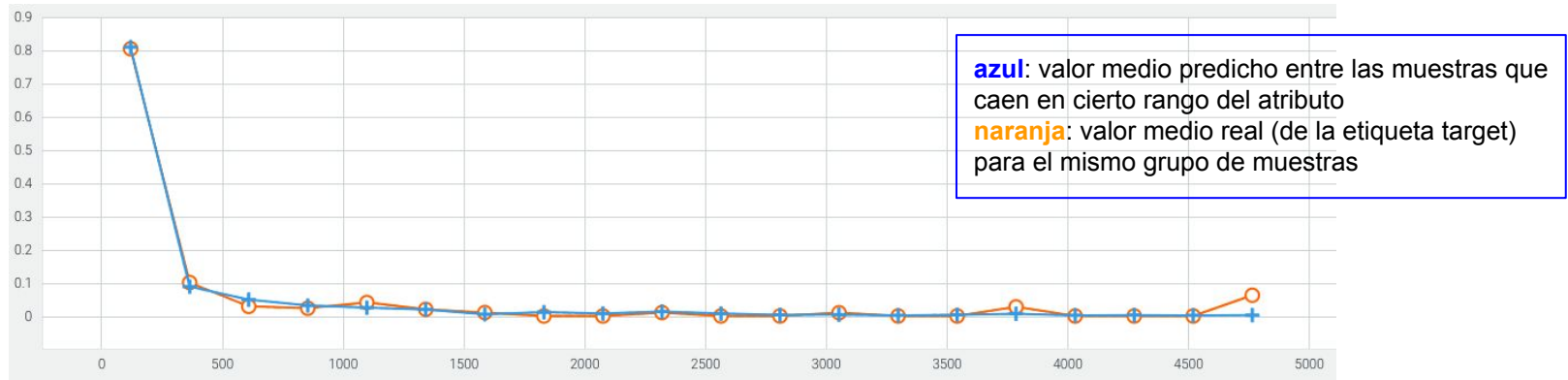
BUENOS DATOS → MAYOR CONFIANZA EN EXPLICABILIDAD A NIVEL PREDICCIÓN INDIVIDUAL

↑ ↓ Prediction: 0.9		
IMPACT	FEATURE	VALUE
+++	atributo_1	10
+++	atributo_2	0
+++	atributo_3	0
+++	atributo_4	0
+++	atributo_5	MISSING
+++	atributo_6	0
+++	atributo_7	0

SHAPLEY values: nos ayudan a interpretar el **aporte** que cada **atributo** tiene en una **predicción** → importante tener bien informados y **fiables** los **valores** con los que explicamos dichas predicciones (decir que el 5º atributo tiene cierta importancia en esta predicción debido a su valor “missing” no aporta conocimiento de negocio)

BUENOS DATOS → MAYOR CONFIANZA EN INTERPRETACIÓN DE MODELOS

- buenos datos → aportan confianza en analizar cómo se comporta mi modelo prediciendo en cada rango de valores del atributo considerado



- Business decisions → además del score del modelo se pueden añadir reglas basadas en los valores de otros atributos considerados por negocio → ejemplo: alto score probabilístico de propensión a fuga + “¿ha realizado cese de domiciliaciones el último mes?”

CONCLUSIONES:

- un proyecto de ciencia de datos requiere una definición precisa del objetivo
- analizar, comprender, procesar y mantener los datos analizados es crucial tanto para métricas globales de la organización como para la generación de modelos que usen dichos datos
- un buen linaje de datos asegura ser capaz de responder ante posibles errores y poder reproducir experimentados de modelado
- preferimos BUEN dato a MUCHO dato
- la calidad de nuestros datos influirá en la interpretación que seamos capaces de dar a la salida de nuestro modelos y traducción en términos comprensibles para negocio

Referencias:

- <https://www.gartner.com/smarterwithgartner/how-to-stop-data-quality-undermining-your-business>
- <https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards>
- https://cloud.google.com/architecture/ml-modeling-monitoring-identifying-training-server-skew-with-novelty-detection#two-sample_statistical_tests
- <https://datascience.stackexchange.com/questions/82301/can-a-novelty-detection-model-overfit/82304#82304>
- https://www.lexjansen.com/wuss/2017/47_Final_Paper_PDF.pdf
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RANSACRegressor.html?highlight=ransac#sklearn.linear_model.RANSACRegressor