COL·LEGI D'ACTUARIS DE CATALUNYA

CONFERENCIA

Barcelona, 29 de junio de 2023

¿Cómo diseñar un sistema de Machine Learning?

Sr. Juan José Nieto, de Glovo

Sr. Heber Trujillo, de Glovo

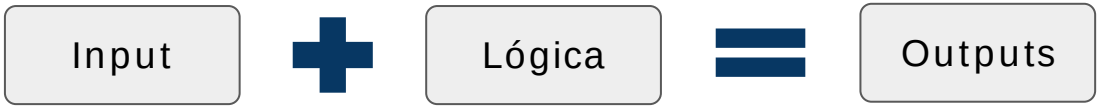
Sistemas de Machine Learning

Juanjo Nieto
Data Scientist @ Glovo

Heber Trujillo
Sr. Data Scientist @ Glovo

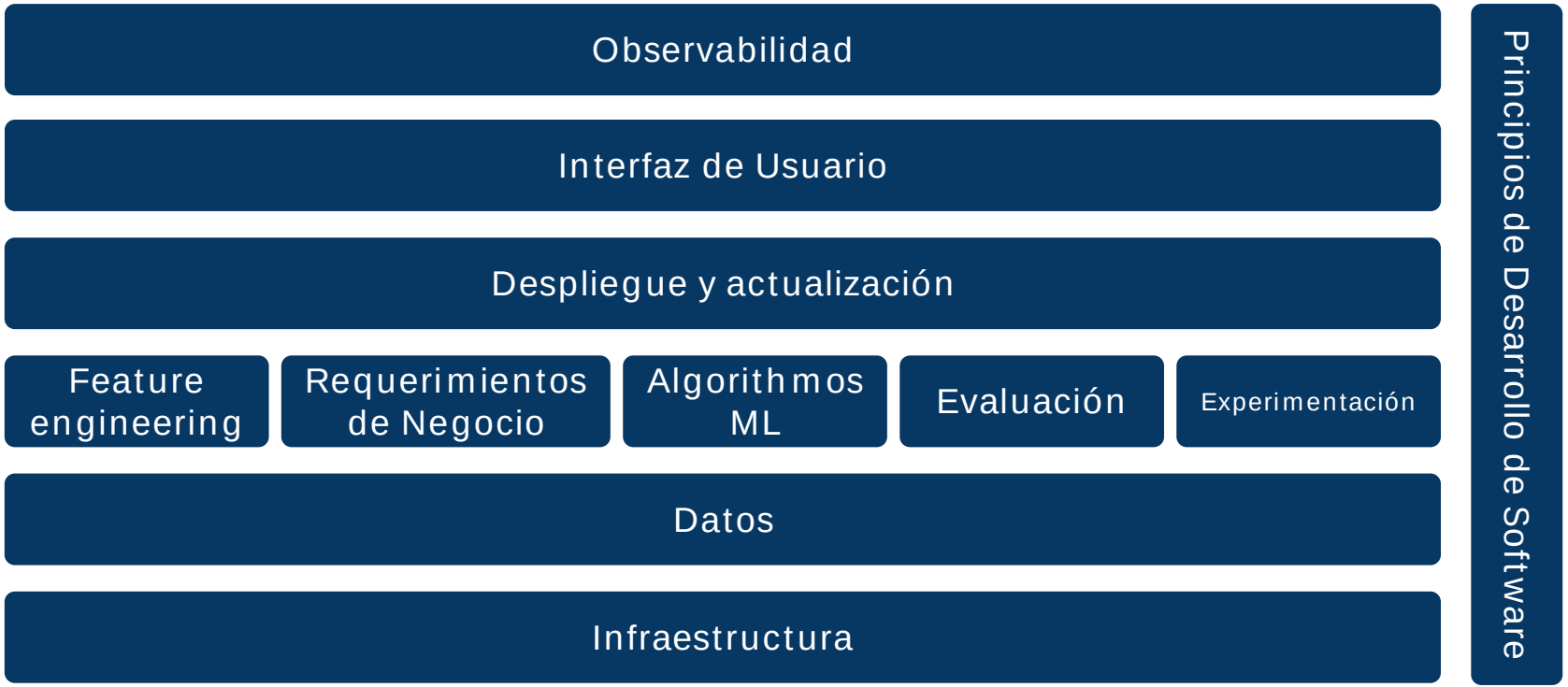
Introducción

Paradigma Clásico



Paradigma ML

Aprender a partir de datos existentes patrones complejos, y utilizar estos patrones para hacer predicciones a escala sobre datos no vistos.


Observabilidad
Interfaz de Usuario
Despliegue y actualización
Feature engineering
Requerimientos de Negocio
Algorithmos ML
Evaluación
Experimentación
Datos
Infraestructura
Principios de Desarrollo de Software

Requerimientos de Negocio

Milton Friedman, "A Friedman Doctrine—The Social Responsibility of Business Is to Increase Its Profits"

Conversión de Clientes

Optimización de Costos

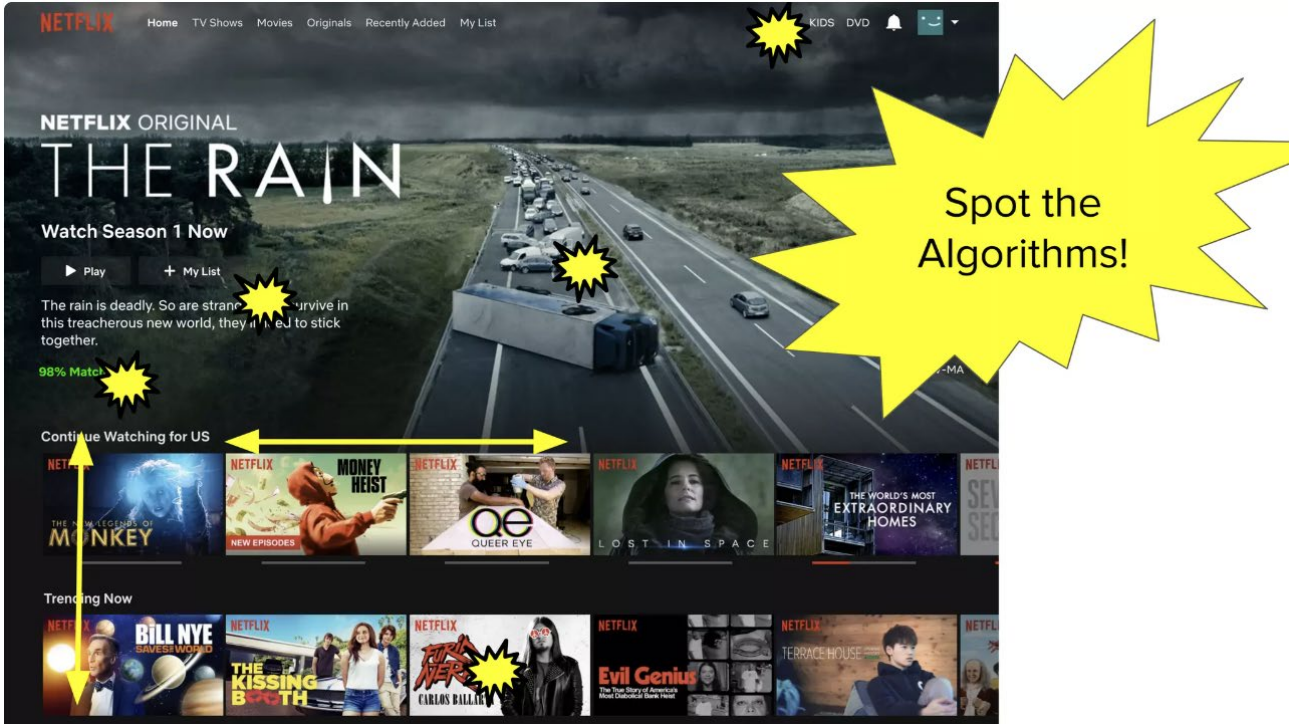
Retención de Clientes

¿Debemos crear un modelo para predecir el abandono (churn) de los clientes?

Ashok Chandrashekar, Fernando Amat, Justin Basilico, and Tony Jebara, "Artwork Personalization at Netflix"

Mejorar la experiencia de nuestros usuarios es clave para aumentar la retención

Ashok Chandrashekar, Fernando Amat, Justin Basilico, and Tony Jebara, “Artwork Personalization at Netflix”



Ashok Chandrashekar, Fernando Amat, Justin Basilico, and Tony Jebara, "Artwork Personalization at Netflix"

$$take\_rate = \frac{\sum reproducciones}{\sum recomendaciones}$$

Mayor take-rate ⇒ Mayor número de horas de Reproducción ⇒ Menor número de cancelaciones

Carlos A. Gomez-Uribe and Neil Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation,"


Ingeniería de Datos

Recolección

Almacenamiento

Acceso

Usuarios

- texto , imágenes, archivos, etc.
- Requiere de muchas validación.

Systemas

- Interacción usuarios, logs, eventos, etc.
- Requiere validaciones.

Bases Internas

- Inventarios, relaciones con clientes, etc.
- No suele requerir de muchas validaciones

Formato

- legibilidad humana: texto vs binario
- Patrones de acceso: filas o columnas

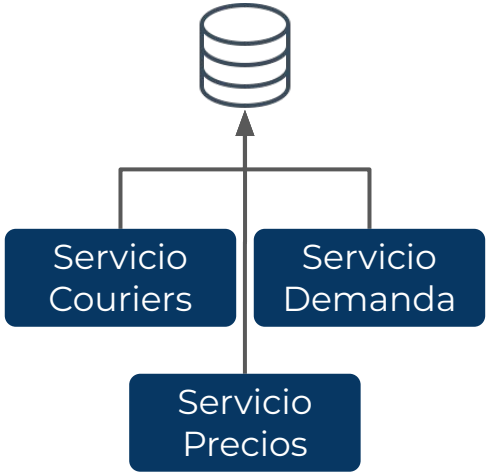
Modelo

- Atributos del objeto a representar
- Datos estructurados o no estructurados

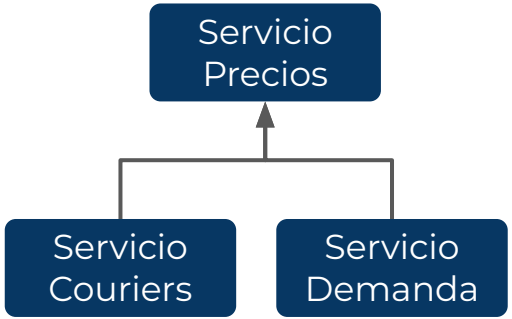
Motor

- Transacional (ACID) o analítico.
- Base de datos Relacional o NoSQL

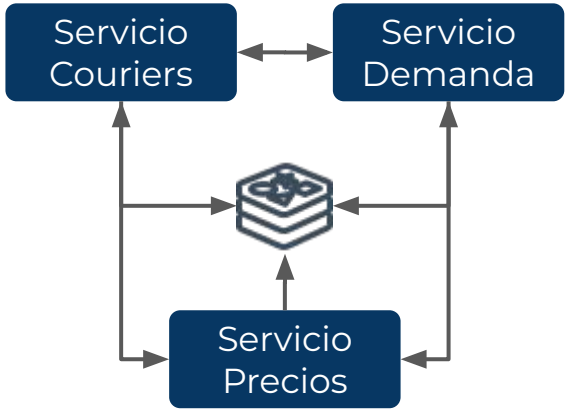
Database



API



Real - Time



Proceso Batch

Proceso Online


Datos

Muestreo

Etiquetas

Desequilibrio

Seleccionar una **subconjunto representativo** de un grupo grande **para deducir información** sobre éste sin tener que estudiar a todos sus miembros.

De la población total para entrenar modelos.

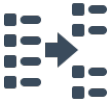
No tenemos acceso a los datos de toda la población.

Usar todos los datos disponibles es inviable por su tamaño.

Para entender el efecto del tamaño de muestra en el rendimiento del modelo

Seleccionar una **subconjunto representativo** de un grupo grande **para deducir información** sobre éste sin tener que estudiar a todos sus miembros.

De la población total para entrenar modelos.



Para crear Train / Test / validation sets.

De todos los eventos para validar nuestro sistema ML.

No Probabilísticos

Muestreo de conveniencia : Seleccionamos todos los datos disponibles.

Snowball: Muestras iniciales nos ayudan a seleccionar las siguientes muestras.

Quota: Selección de muestras para alcanzar una proporción de estratos arbitraria.

Probabilísticos

No Probabilísticos

Probabilísticos

Pros:

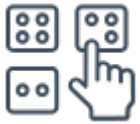
Rápido y fácil de implementar

Agiliza la construcción del sistema ML

Cons:

No representativo de la realidad

Introducen sesgos de selección a los modelos



No Probabilísticos

Pros:

Rápido y fácil de implementar

Agiliza la construcción del sistema ML

Cons:

No representativo de la realidad

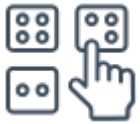
Introducen sesgos de selección a los modelos

Probabilísticos

Aleatorio Simple: Selección de muestra usando una distribución uniforme

Pros: Fácil de implementar

Cons: Categorías raras pueden no estar presentes en la muestra.



No Probabilísticos

Pros:

Rápido y fácil de implementar

Agiliza la construcción del sistema ML

Cons:

No representativo de la realidad

Introducen sesgos de selección a los modelos

Probabilísticos

Estratificado: Seleccionar la muestra desde subgrupos usando una distribución uniforme.

Pros: Categorías raras estarán representadas

Cons: Introduce sesgos en inferencia

No Probabilísticos

Pros:

Rápido y fácil de implementar

Agiliza la construcción del sistema ML

Cons:

No representativo de la realidad

Introducen sesgos de selección a los modelos

Probabilísticos

Weighted Sampling, Reservoir Sampling, Importance Sampling, etc …

Valores o categorías asignados a los datos que representan la **respuesta** deseada que el **modelo** debe **aprender a predecir**.

Manuales

Lento, costoso, un problema con la privacidad.

Multiplicidad de etiquetas, diferentes niveles de calidad.

Auto-generadas

Escalable, flexible, un problema con la privacidad.

Requiere una estrategia para gestionar feedback loops, e.g.

Diferencia substancial en el número de muestras de las **clases** dentro de nuestros datos, e.g., detección de fraude.

Desafíos

Errores asimétricos, peor fallar en clases minoritarias.

Modelo sub-óptimo, aprende un patrón trivial.

Señal débil, imposible detectar clase minoritaria.

¿Por qué ocurren?

Error durante la recolección de datos

Sesgo en el muestreo de los datos, e.g. email spam.

Inherente, al problema que buscamos resolver.

Alternativas

Métrica, evaluar el modelo ponderando las clases.

sub/over-sampling, modificar la distribución de los datos.

Algoritmo, cost-sensitive learning, class balance / focal loss.


Feature Engineering

Características de los datos que nuestro modelo usará para **aprender patrones ocultos**.

Engineered Features

Características **diseñadas manualmente** usando conocimiento específico del problema a resolver.

Learned Features

Características **automáticamente extraídas por el modelo**, e.g. DL - NLP no suele requerir lemmatization

Operaciones

Data Leakage

Evaluación

Ausencia de valores en un conjunto de datos, puede **complicar** el **análisis** y la **interpretación** de los **resultados**.

Sistemática y relacionada con la propia variable estudiada.

Sistemática y relacionada con otra variable.

Completamente aleatorio.

Convertir variables **categóricas**, como el nombre de las provincias, en una forma **numérica** comprensible para los algoritmos de ML.

La alta cardinalidad puede introducir sesgos si decidimos agrupar.

Las categorías cambian con respecto al tiempo, cómo gestionar nuevas categorías.

Una forma de **la variable objetivo se filtra** en el conjunto de **features** que utilizamos para **entrenar el modelo**, y esta información **no es accesible** durante **inferencia**.

Ignorar la correlación temporal al momento de crear y evaluar un modelo, e.g. stock prices.

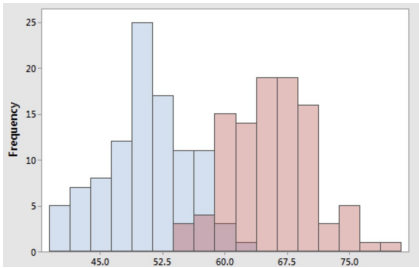
Utilizar todos los datos para transformar las features, e.g. escalar con estadísticas globales.

Mala gestión de los datos duplicados, e.g. oversampling antes de dividir los datos.

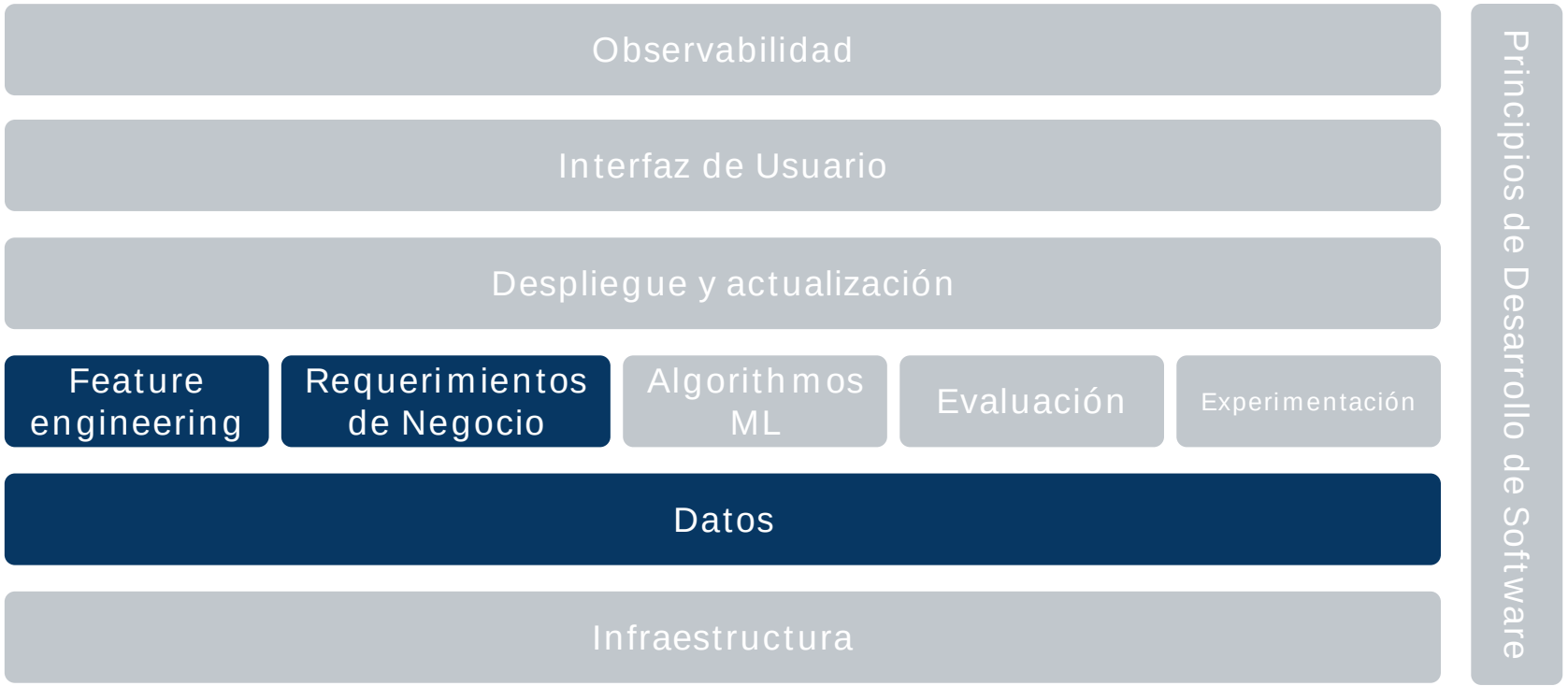
Añadir **más features** por lo general **mejora el rendimiento** de un **modelo** ML, pero **no** necesariamente **mejora** el rendimiento del **sistema ML**, e.g. aumento de latencia y recursos, probabilidad de data Leakage, overfitting, etc.

Importancia de una feature, cuánto se deteriora el rendimiento de ese modelo si eliminamos una feature.

Generalización de una feature, el rango de valores de la feature debe ser similar en inferencia.



Conclusiones


Observabilidad
Interfaz de Usuario
Despliegue y actualización
Feature engineering
Requerimientos de Negocio
Algorithmos ML
Evaluación
Experimentación
Datos
Infraestructura
Principios de Desarrollo de Software


Q & A

Apéndice

Para entender cómo influyen las métricas de un modelo ML en los objetivos de negocio, es necesario realizar experimentos, e.g. A/B test.

Método	Descripción	¿Requiere etiquetas?
Weak Supervision	Heurísticas simples, labeling functions	No, pero es recomendable tener algunas para guiar el desarrollo de las heurísticas.
Semi-Supervision	Usa la estructura de los datos para aumentar el número de muestras etiquetadas.	Sí, un pequeño número que servirá de base para generar más.
Active Learning	Seleccionar las muestras más útiles y etiquetarlas.	Sí, entre más mejor

	Entorno Académico	Entorno de Producción
Requerimientos	Mejorar el rendimiento tanto como sea posible.	Diferentes stakeholders tienen diferentes requisitos.
Datos	Conjuntos de datos de referencia estáticos.	Dinámicos, feedback loops.
Interpretabilidad	No suele ser considerado.	Debe ser considerado.

COL·LEGI
D'ACTUARIS
DE CATALUNYA

actuaris@actuaris.org
www.actuaris.org