Big Data + IA + Finanzas

(No tan) Nuevas maneras de usar tecnología al procesar información financiera

Wanda Rincón Cadena

Sobre mí...

- Ingeniera de Sistemas UIS
- Msc Matemática Aplicada
- Trabajo como Data Analyst en MO Tecnologías
- Ex- CUSOL-UIS, colaboradora en Django Girls Colombia y Python Bogotá

¿Qué es una Fintech?

Muchas experiencias con los bancos eran en extremo desagradables.

- Filas, procesos lentos.
- Mucha gente no estaba bancarizada o no puede obtener un crédito de entidades bancarias (inclusión financiera).

Fintech

> *Una empresa que se dedica a innovar con tecnologías para soluciones financieras (p2p lending, crowdfunding, pagos digitales, etc.)*

Fintech en Colombia

Los Gobiernos y reguladores han entendido el poder que tiene la tecnología en la creación de un sistema financiero más inclusivo, **luchar contra la pobreza e impulsar la economía.**

> *Colombia es el tercer país de América Latina con el mayor ecosistema Fintech, solo superado por Brasil y México.*

Cómo están cambiando la economía ?

Muchas fintech están cambiando el sector bancario y creando oportunidades.

Se necesita un **equipo**:

- Ingenieros de datos
- Arquitectos de Big data
- **Científicos de datos (antes BI)**
- Economistas, estadísticos, matemáticos..
- Desarrolladores
- Gente interdisciplinar !

Análisis de Datos

Las empresas suelen contar con analistas de negocios, que generan reportes, gráficas, histogramas para un mejor entendimiento de los datos que se poseen.

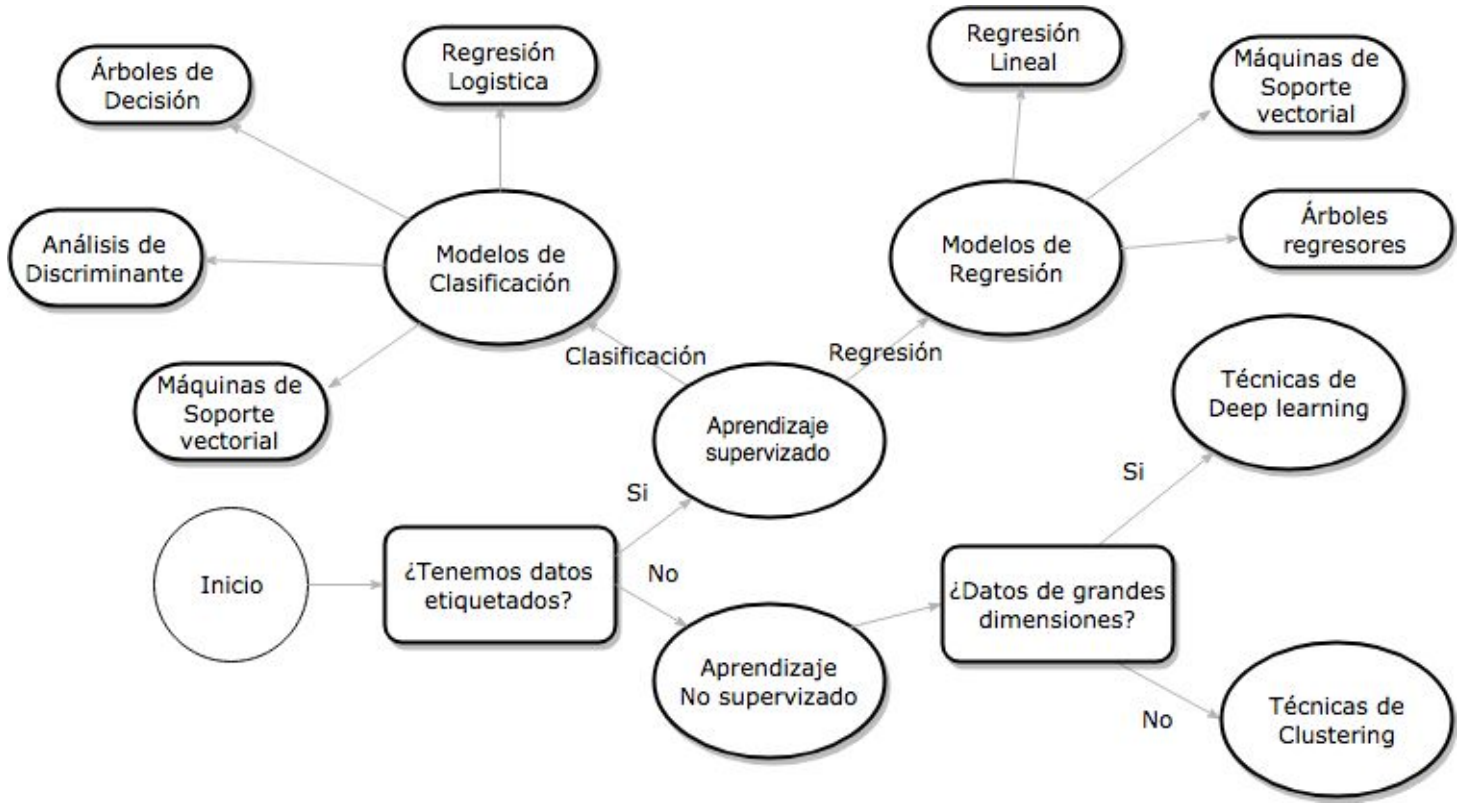
¿Qué es machine learning?

Machine learning es un conjunto de modelos matemáticos que pueden **descubrir patrones** o relaciones relevantes entre datos permitiendo obtener **nueva** información.

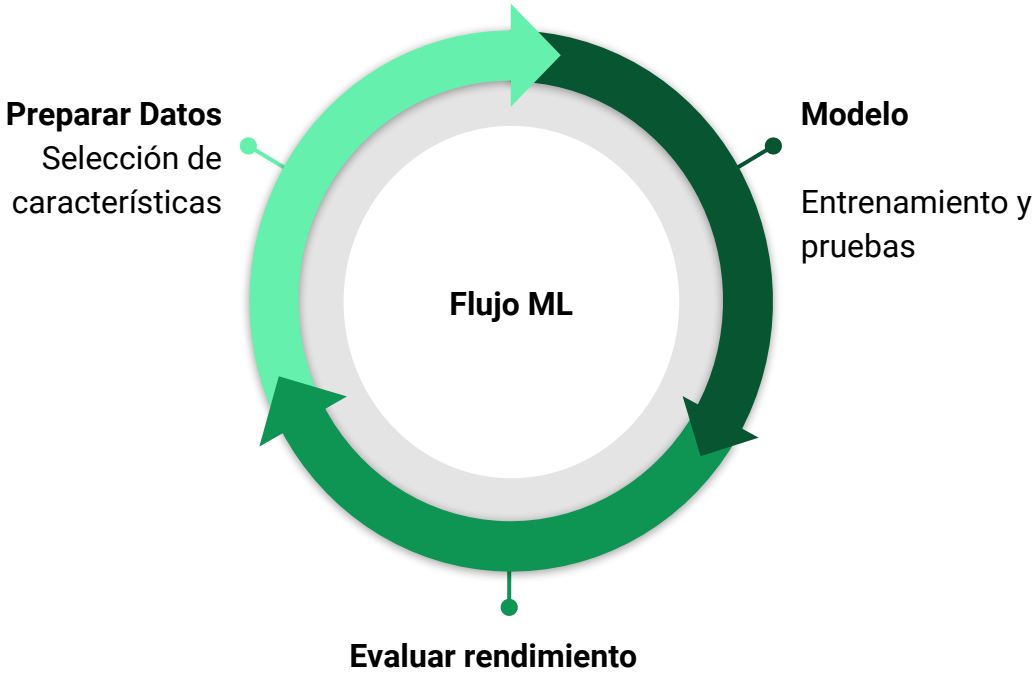
"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." Mitchell, T. (1997).

Tecnologías Machine Learning

OBJ



Flujo de trabajo en ML



Seguir “piloto” y repetir ciclo hasta poder llevar a producción!

Flujo de trabajo en ML

- Paso 0: Tener buenos datos.

  (Garbage IN, Garbage OUT):

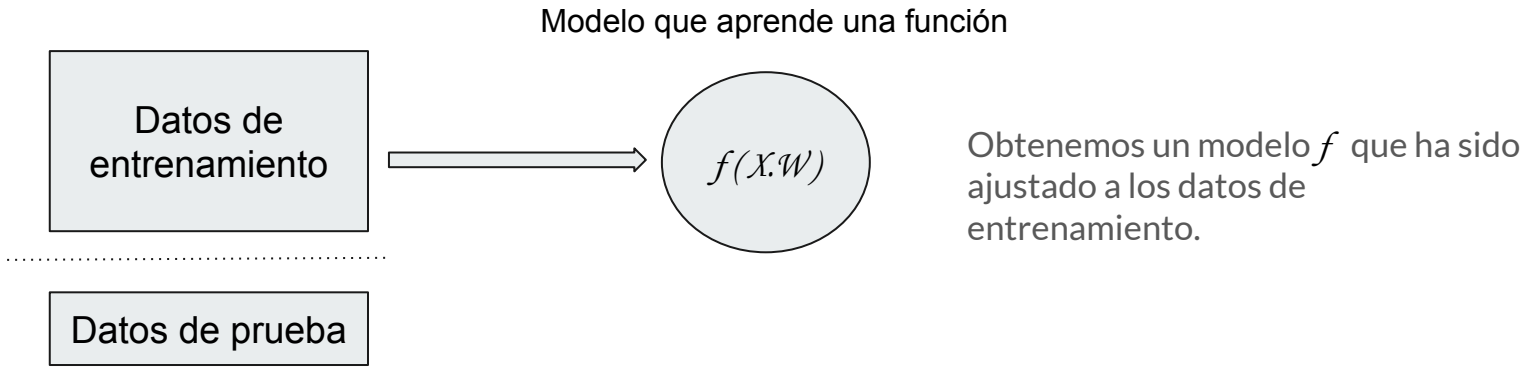
  - ★ Decidir qué hacer con los valores atípicos.
  - ★ Filtrar datos ruidosos.
  - ★ Remover errores de los datos.
  - ★ Decidir qué hacer con datos incompletos.

Flujo de trabajo en ML

- Determinar el **tipo de tarea** (forecasting, clustering, regresión o clasificación)
- Extracción de **características**
- Selección de características
- Exploración de los datos y estadística, graficación.
- **Entrenamiento** y ajuste de modelos
- Selección del **mejor modelo**
- **Prueba** del modelo
- Despliegue!

Entrenamiento

En esta etapa **uno o varios** algoritmos pueden ser usados para resolver el problema, los datos disponibles son divididos en conjuntos de entrenamiento y prueba (datos no vistos).



Prueba

Finalmente, el mejor modelo es probado en datos no vistos durante el entrenamiento (si hay suficientes datos). El modelo es modificado si se encuentran problemas comunes como lo son el sobreajuste.

Después de las pruebas el modelo es desplegado y re-entrenado según la necesidad.

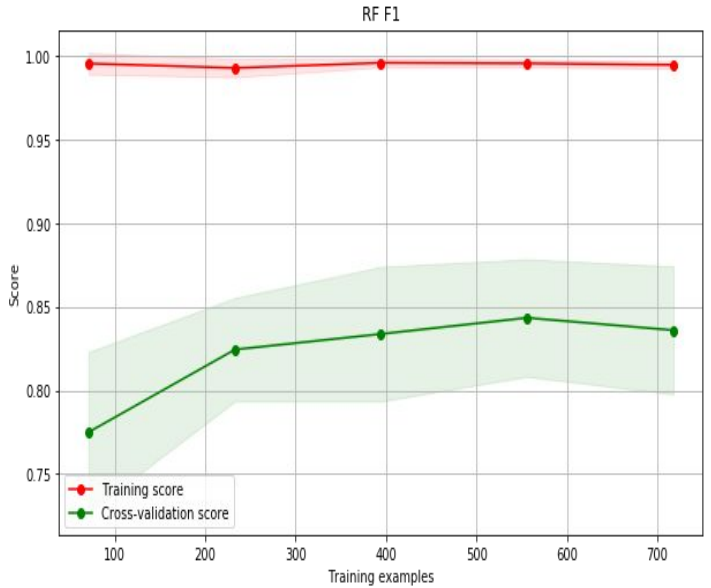
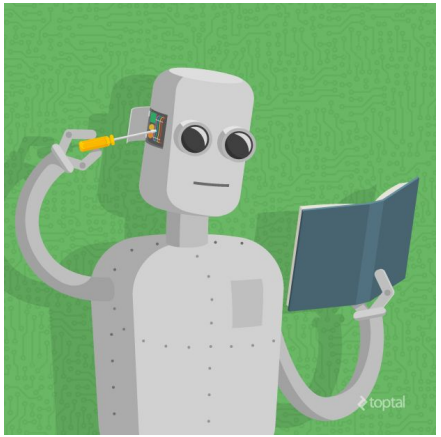


Figura: Curva de aprendizaje. Describe cómo se comporta el algoritmo con el aumento de datos (experiencia) en entrenamiento y prueba.

Problemas que suelo resolver usando ML

- ➔ Calcular credit score (algo así como data crédito)
- ➔ Probabilidad de default (riesgo)
- ➔ Perfiles de clientes
- ➔ Forecasting de transacciones para clientes



Ejemplo 1: Calcular credit score

Problema: Darle un puntaje a un cliente según qué tan bueno es pagando sus deudas, basándonos no en datos personales sino en sus hábitos transaccionales, ¿como podemos hacer esto?

La manera tradicional consiste en darle un peso a mano a diferentes comportamientos y sumarlo en un “score” que cuanto más alto, menos riesgo representa para el prestamista.

Datos:

x = [x1, x2] → Variables predictoras

Y → Variable a modelar. Es continua (no es una categoría)

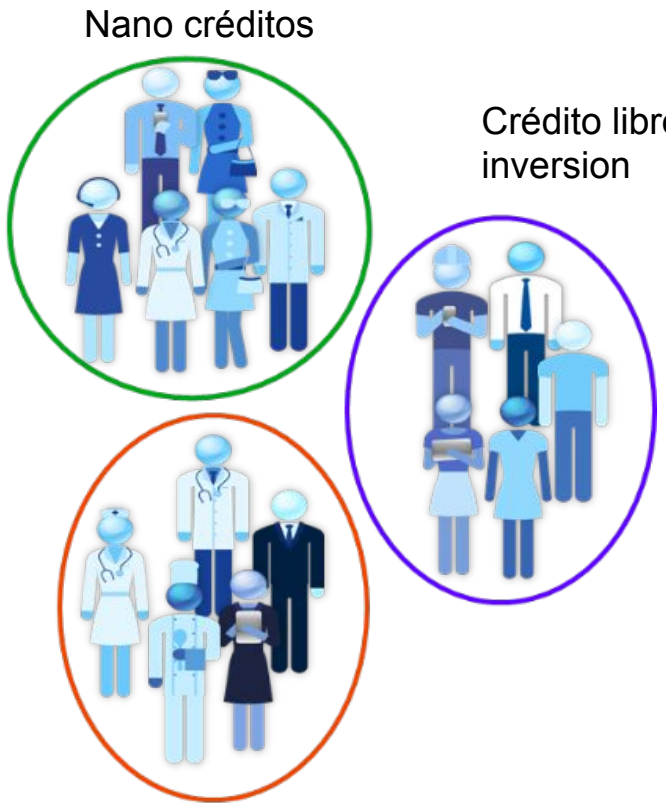
Podríamos plantear esto como un modelo de **regresión!**

Id	x1	x2	Y
id1	34	87	97
id2	12	3	20
id3	23	89	56

Ejemplo 2: Conocer nuestros clientes

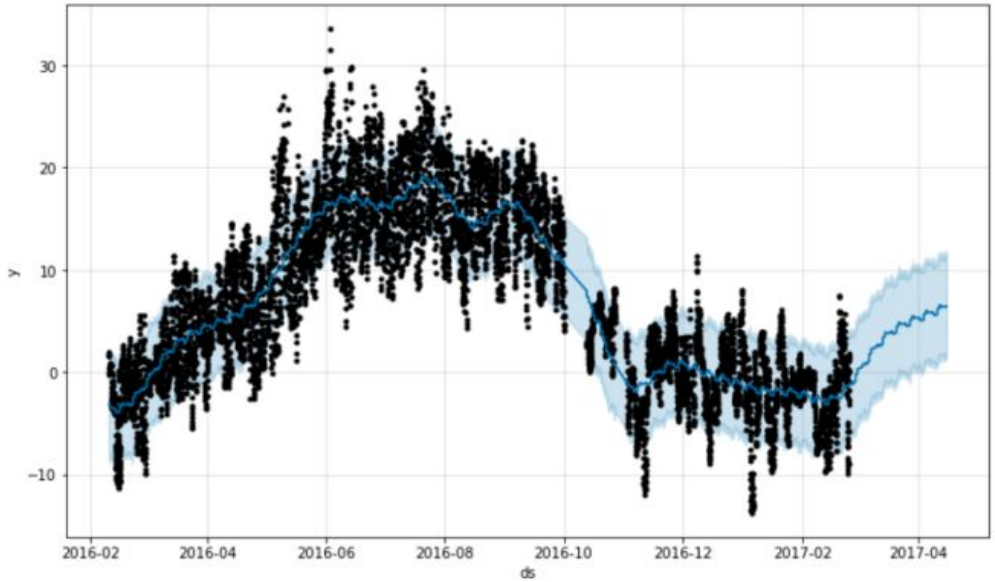
Problema: Tenemos datos personales, transaccionales y de otros tipos de clientes, no poseemos una categoría ni una variable a predecir. Queremos conocer más estos clientes para ofrecer créditos a su medida.

Como esto no es un problema de aprendizaje supervisado podríamos intentar clusterización !


Nano créditos
Crédito libre inversion
Crédito para vivienda

Ejemplo 3: Predecir comportamiento de un cliente

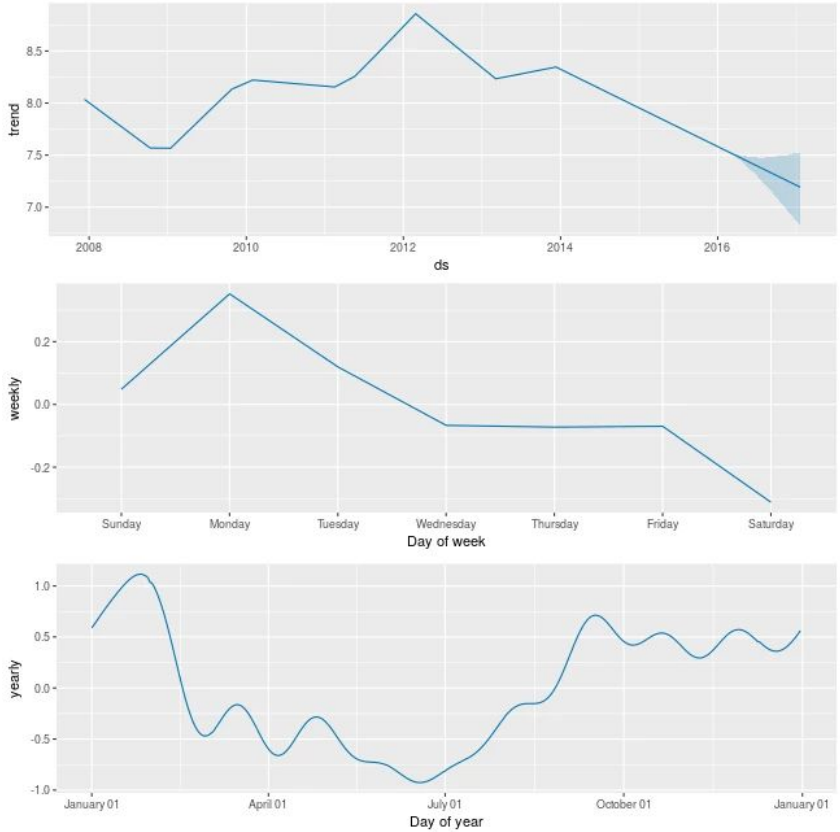
Problema: Tener un estimado de las transacciones o ventas de un tendero, predecir si hará fraude.



Las series de tiempo suelen ser muy ruidosas, tener componentes de estacionalidad que no dejan apreciar a simple vista el comportamiento en el tiempo.

Los métodos usados sobre secuencias temporales incluyen:

- ARIMA
- Descomposición de series (tendencia, estacionalidad, ruido)
- Deep learning (CNN, RNN, LSTM etc)



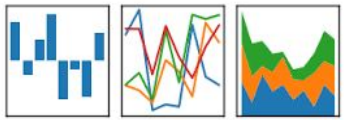
Data...

Con el desafío de que dependiendo del problema podemos tener pocos o muchos datos!

- **Pocos datos** → ¿ Cómo conseguir una buena predicción?
- **Muchos datos** (big data) → ¿ Cómo procesar estos datos de manera eficiente?

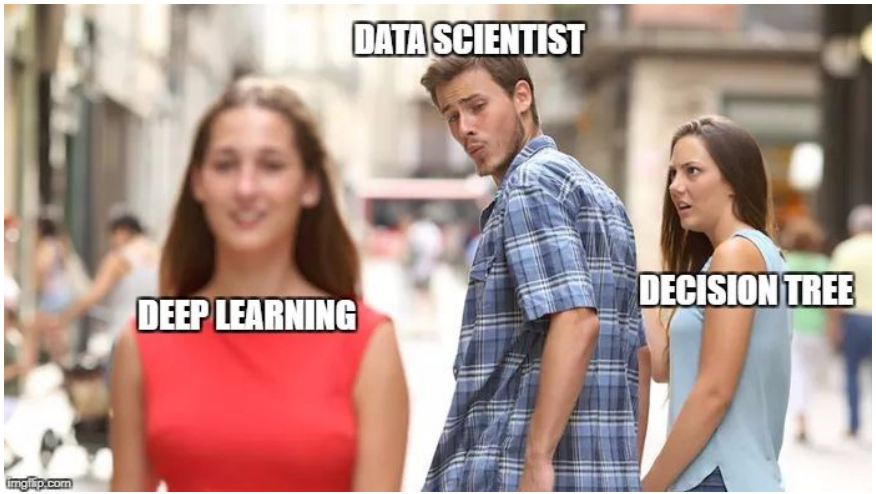
Algunos consejos prácticos

- Visualizar siempre! ANTES de aplicar modelos, la exploración es clave para empezar a tratar de entender los datos.
- Librerias recomendadas (data scientist starter kit!):



Algunos consejos prácticos

- Empezar con un modelo sencillo … hacer el ejercicio de interpretar un poco antes de saltar a un modelo mucho más complejo y costoso!



Preguntas

En github estoy como @wandarinca

https://www.linkedin.com/in/wanda-rincon-83549098/