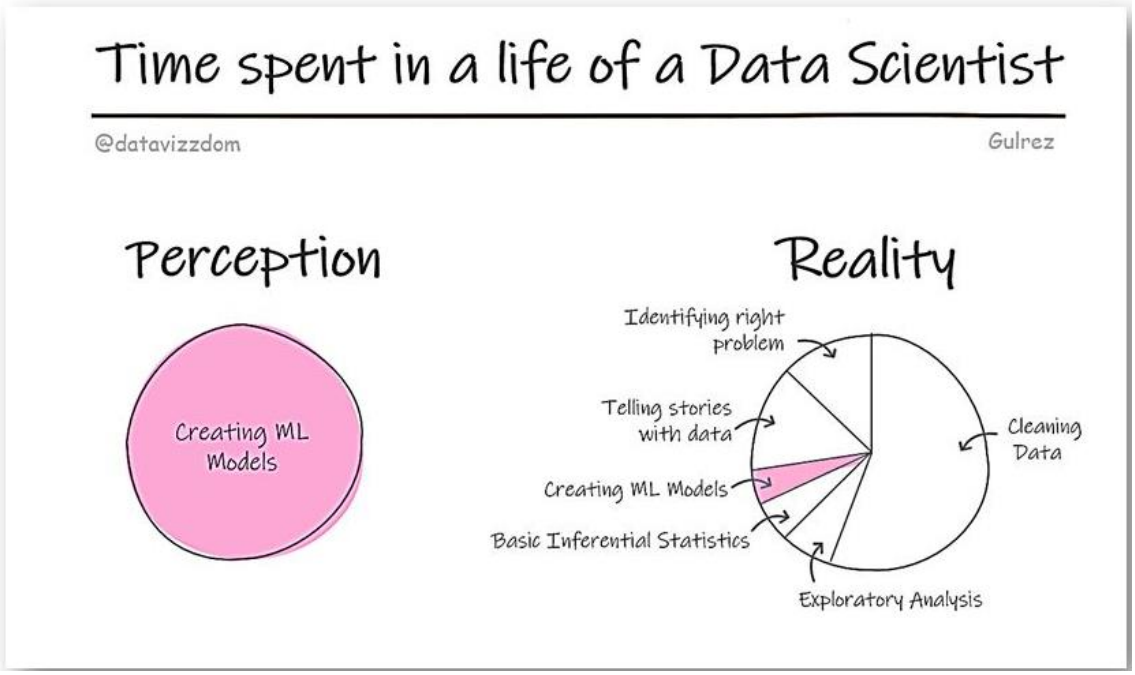


**Workflow de un proyecto completo de aprendizaje supervisado de Data Science:**

- Aspectos de Ingeniería y Gestión:
- Definición de objetivos y necesidades ¿se precisa más velocidad que precisión?
- Limitaciones hardware, escalabilidad, etc.
- Complementación cloud, codificación para producción, entorno de desarrollo, etc

(...)

---

**Recuerda que en una empresa normalmente trabajamos en un equipo multidisciplinar. Según las necesidades de los otros departamentos de una empresa, cada punto puede extenderse con más necesidades específicas.**

Es muy importante el tener en cuenta que si eligimos una semilla (seed), puede que nuestro algorimo no obtenga el resultado óptimo. Esto se debe a que cogerá partes de datos específicas y se harán reordenaciones específicas.

1. ¿Existen datos? Si no tenemos datos, el primer objetivo es conseguirlos.
2. ¿Están etiquetados? si no, lo hacemos nosotros a mano (hay software que podría ayudarnos).
   - Si no podemos hacerlo nosotros ni el software, entonces se trata de un problema no supervisado.
   - Si hay labels pero no en todos los datos, entonces se trata de un problema semi-supervisado.
3. Una vez tengamos los datos etiquetados, debemos empezar nuestra Exploración y Análisis de Datos(EDA):

- Sacamos nuestra "X" e "y" de los datos.
- Si tenemos datos categóricos, pasaremos un encoder a las columnas no numéricas para realizar la transformación.
- Realizamos la matriz de correlación y otras gráficas para visualizar y entender mejor nuestros datos. Podemos quitar las columnas que tengan menos correlación (cercanas al 0). Esa decisión alterará a la precisión de nuestro modelo ya que entrenará menos datos.
- Podemos realizar una normalización de los datos. Esto es opcional tanto para una como para todas las columnas numéricas (no categóricas).
- Realizamos otras modificaciones de nuestros datos: quitar columnas con un %NaN mayor a N, reemplazar NaN por media, agregar columnas que sean operaciones estadísticas entre columnas (media, mediana, varianza, etc), agrupar por datos categóricos, ...
- Podemos representar de nuevo la matriz de correlación y otras gráficas para ver si con las modificaciones realizadas se modifica y podemos sacar diferentes conclusiones.
- (...)

4. Realizamos el split de datos:

    - Si los datos son demasiados para los recursos hardware que tenemos, podemos realizar un split más pequeño para trabajar con ellos en primera instancia.
    - Obtenemos el conjunto de train y de test.
5. Elegimos el algoritmo a utilizar:

    - Podemos realizar un GridSearch para encontrar el mejor modelo entre varios y las features adecuadas.
        - Si no son muchos datos, podemos usar con el X_train completo.
        - Si tenemos limitaciones, podemos usar el sample generado para ver qué algoritmo se comporta mejor.
        - Aquí vamos a ver un primer score. Sería buena idea volver al punto 3 y realizar más cambios sobre nuestro conjunto de datos.
6. Entrenamos el modelo. Tenemos varias posibilidades:

    6.1. Si nuestro algoritmo no nos permite validación cruzada:

        6.1.1. Usar todos los datos si nuestro hardware lo permite y no necesitamos controlar lo que ocurre durante el entrenamiento.

        6.1.2. Usar el sample de datos más pequeño para ver qué score puede ofrecer. Volver al punto 3 y realizar las modificaciones pertinentes.

6.1.3. Volveremos al punto 3 para realizar más cambios a nuestro dataset si lo vemos necesario.

6.2. Si nuestro algorimo nos permite validación cruzada:

6.2.1. Tenemos las mismas opciones que en la 6.1

6.2.2. Realizar validación cruzada con todos los datos.

6.2.3. Realizar validación cruzada con el sample de datos para ver qué score puede ofrecer. Volver al punto 3 y realizar las modificaciones pertinentes.

6.2.4. Realizar validación cruzada con todos los datos poco a poco (entrenamiento incremental, online, mini_batch, en caliente). Esta sería la opción óptima ya que nos permite controlar lo que va ocurriendo y ver cómo evoluciona el entrenamiento.

7. Sacar score de nuestros datos:

7.1. Si no son los deseados, debemos volver al punto 3, 4 o 5 dependiendo de las necesidades.

7.2. Si lo son, 8.

8. Realizar un entrenamiento completo de los datos. Es decir, ya no se utilizarán los datos del split, no habrá X_test. Aquí volvemos a realizar el análisis con todos los datos usando alguna opción del punto 6 (sin la parte del sample). Una vez realizado, guardamos el modelo con sus estadísticas anotadas.
9. Partes de ingeniería y otras necesidades (...)