DigitalHouse >
Coding School

DATA SCIENCE

UNIDAD 1
MÓDULO 1

Flujo de Trabajo en Data Science

Agosto 2017

Este breve documento tiene por objeto ampliar algunos aspectos que el flujo de trabajo en Data Science. Como se ha mencionado en la clase el esquema propuesto resulta una "estilización" del proceso real de trabajo. Se encontrarán con numerosas iteraciones entre las diversas etapas. Se presentan aquí una reseña de algunas de dichas etapas junto con las "preguntas" y "tareas" más comunes en cada una.

- **IDENTIFICAR EL PROBLEMA: entenderlo**
  - Identificar los objetivos del negocio/producto/problema
  - Plantear y/o Identificar objetivos y criterios para el éxito
  - Formular un conjunto de preguntas que permita identificar el(los) dataset(s) apropiado(s)

- **ADQUIRIR: obtener los datos**
  - Dato ideal versus dato real: es común comenzar un proyecto imaginando el dataset ideal para abordarlo
  - En el proceso de investigación real del problema es habitual encontrar que los datos que tales datasets ideales no existen o son muy costosos de producir. En el mismo proceso deberemos investigar y conocer las limitaciones y potencialidades de los datos reales (disponibles en tiempo y forma).
  - Será necesario decidir si tales limitaciones son tan grandes como para inhabilitar su uso o si pese a las mismas es posible abordar de forma consistente los problemas planteados y generar una respuesta aceptable.
  - ¿Pueden utilizarse dichos datos reales?
  - Preguntas habituales en esta etapa:
    - ¿Cuál es el dataset "correcto"? ¿Cuántos datos hay y cuántos son suficientes?
    - ¿En qué medida logran enfocar el problema/pregunta de forma correcta?
    - ¿Cuál es el grado de confiabilidad de los datos? ¿Cómo fueron recolectados?
    - ¿Cuál es el nivel de agregación de los datos? ¿Es adecuado? ¿Se puede acceder al nivel adecuado?

  - ¿Cómo tendremos que acceder a los datos?
    - Importar datos de la web (Google Analytics, HTML, XML. etc.)
    - Importar datos de un archivo (.CSV, ,XLSX, .TXT, .XML,

.JSON, etc.)
    - Importar datos de una base de datos preexistente (SQL)
    - Montar una estructura nueva local o remota
    - Será necesario identificar las herramientas más apropiadas para trabajar con los datos (de acuerdo a su forma, tamaño, etc.)

- **PARSEAR: entender los datos**
    - Situación muy común: uso de datos secundarios o recolectados previamente
    - Importante: estudiar y conocer todos los "metadatos": diccionarios, libros de códigos, documentos metodológicos, etc.
    - ¿Por qué sería importante conocer la forma en que fueron recolectados los datos?
    - Diccionario o libro de códigos: documentos que explican qué son nuestros datos y cómo están formateados. Por ejemplo:
        - **Variable | Description | Type of Variable**
          Profession | Title of the account owner | Categorical
          Company Size | 1- small, 2- medium, 3- large | Categorical
          Location | Planet of the company | Categorical
          Days Since Last Delivery | Integer | Continuous
          Number of Deliveries | Integer | Continuous
    - Muchas veces se nos provee datos secundarios o datos que fueron recolectados previamente. En estos casos podemos (y es sumamente necesario hacerlo) estudiar y conocer todos los "metadatos": diccionarios, libros de códigos, documentos metodológicos, etc. que nos brindarán información acerca de la forma en que fueron recolectados los datos.
    - Tareas comunes en esta etapa:
        - Leer la documentación provista con los datos (por ejemplo, los diccionarios como el mostrado previamente)
        - Realizar un análisis superficial a través de filtros, ordenado, visualizaciones básicas
        - Describir la estructura de los datos y la información recolectada
        - Explorar variables, tipos de datos, etc.
        - Realizar un primer análisis de tendencia, outliers, etc.

- Chequear la calidad y la integridad de los datos

- **MINAR: preparar, estructurar y limpiar los datos**
  - Con frecuencia (casi siempre), nuestros datos deberán ser "limpiados" previamente a la realización del análisis
  - ¿Qué significa limpiar los datasets? ¿Por qué es necesario?
  - Algunas tareas comunes en esta etapa incluyen:
    - Muestrear los datos, determinar la estrategia de muestreo necesaria
    - Iterar y explorar valores extremos (ouliers) y valores perdidos o NULL
    - Revisar datos cualitativos versus cuantitativos
    - Formatear y limpiar los datos (por ejemplo, fechas, signos numéricos, formatos)
    - Definir las formas apropiadas de lidiar con los valores perdidos (cleaning, imputación, exclusión, etc.)
    - Recategorización, manipulación, "slicing", formateo, integración de datos
    - Determinar los métodos más apropiados de limpieza y agregación de la información
    - Crear nuevas columnas (variables) derivadas de los datos: recodificaciones, información nueva, etc.

- **REFINAR: Análisis Exploratorio de Datos (AED) e Iteración**
  - AED - o análisis exploratorio de datos-. Por ejemplo, algunas estadísticas básicas, observar la media, analizar frecuencias de los datos. Ejemplo:
- **Variable	Mean (STD) or Frequency (%)**
Number of Deliveries	50.0 (10)
NYC	50 (10%)
LA 9	100 (20%)
Portland	100 (20%)
Seattle 8| 100 (20%)
Other	150 (30%)
  - Tales estadísticos descriptivos permiten:
    - Identificar tendencias y outliers
    - Decidir cómo lidiar con los outliers (exclusión, filtro, imputación, etc.)
    - Calcular estadísticas descriptivas e inferenciales
    - Construir algunas visualizaciones preliminares
    - Documentar hallazgos preliminares
    - Transformar los datos

- **CONSTRUIR: crear un modelo sobre los datos**
  - Construir modelos predictivos basados en los datos. Por ejemplo, una afirmación basada en un modelo sería la siguiente:
    - "Completamos una regresión logística usando Statsmodels. Calculamos la probabilidad de que un cliente realice un pedido a la empresa..."
  - Determinar la probabilidad de que un cliente realice una compra => un problema de clasificación
  - Algunos de los pasos para construir modelos:
    - Seleccionar el modelo apropiado
    - Estimar el modelo
    - Testear y entrenar el modelo
    - Evaluar y mejorar el modelo

  - Una vez que hemos limpiado y explorado los datos, es tiempo de empezar a construir modelos predictivos basados en las variables dependientes que nos interesan o en los supuestos del modelo que estamos usando. Un ejemplo de una afirmación basada en un modelo sería la siguiente::

- **PRESENTAR: comunicar los resultados del análisis**
  - Es una parte crítica del análisis
  - Si no se comunican de forma clara y concisa los hallazgos puede suceder que NO SE USEN
  - Al menos una oración simple y descriptiva de los resultados
    - "Los clientes de gran tamaño tienen el doble (CI 1.9, 2.1) de chances de realizar un pedido que aquellas firmas de menor tamaño."
  - ¿Qué pensas que significa el CI? ¿Por qué deberíamos incluir esta información entre los hallazgos?
  - Las presentaciones de Data Science puede ser MUCHO más complejas y estimulantes (ver por ejemplo, el blog de Nate Silver: 538).
  - Fundamental pensar en la audiencia:
    - anticipar preguntas
    - “testear” la presentación con algunas personas.
  - Presentación para colegas Data Scientists será muy diferente a
    - una orientada a ejecutivos que deben tomar una decisión de negocio
    - o a funcionarios que deben implementar una política pública.

- No importa qué tan bueno, brillante o innovador es el modelo o qué tan “iluminadores” son los hallazgos que produce... si no se comunican de forma efectiva y comprensible puede suceder que NO SE USEN

- Algunas claves para una buena presentación:
  - Resumir hallazgos con alguna narrativa/historia
  - Refinar las visualizaciones para una mejor comprensión
  - Presentar tanto los supuestos como las limitaciones de análisis
  - Determinar la integridad y validez del análisis
  - Testear y evaluar la efectividad de la presentación previamente

- **ALGUNOS COMENTARIOS SOBRE “ITERACIONES”**
  - Necesario repetir tareas, reformular objetivos para poder entender mejor los datos, clarificar los modelos y refinar las presentaciones.
  - Por ejemplo, luego de presentar tus hallazgos, podrías:
    - Identificar nuevos problemas y preguntas para futuros análisis
    - Crear un resumen visualmente efectivo
    - Considerar las diferentes necesidades de los asistentes/ejecutivos/tomadores de decisiones y cómo el reporte podría adaptarse a sus necesidades
    - Identificar las limitaciones del análisis

  - Una vez más: el flujo de trabajo en Data Science no es lineal, puede bifurcarse y plegarse sobre sí muchas veces a lo largo de un proyecto.
- En cualquier parte del proceso puede ser necesario repetir tareas, reformular objetivos para poder entender mejor los datos,clarificar los modelos y refinar las presentaciones.