II. The Data Science Process

Introduction to Data Science

Patricio Mallea[1]

[1]Data Scientist

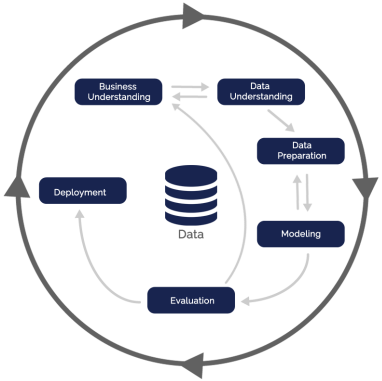
May 11, 2020

Overview I

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

The Data Science Procces: Un acercamiento

DataOwl



La metodología **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*) es un proceso de ciclo de vida para un proyecto de datos, en donde subdivide el proyecto en 6 fases. No obstante, la secuencia no es estricta, de hecho, la mayoría de los proyectos avanzan y retroceden entre fases según sea necesario.

1. Business Understanding

1. Business Understanding

Esta fase inicial se enfoca en comprender los objetivos y requisitos del proyecto desde una perspectiva comercial o de generación de valor, luego convertir este conocimiento en una definición de problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.

1. Determinar los objetivos del proyecto.
2. Evaluar el contexto.
3. Determinar las metas de Data-Mining.
4. Generar el plan del proyecto.

Los resultados de esta fase generalmente son resumidos en un reporte corto, y opcionalmente las documentaciones requeridas para las herramientas a utilizar.

2. Data Understanding

2. Data Understanding

Luego viene el entendimiento de los datos, que comienza con la recopilación inicial y continúa con actividades que le permiten familiarizarse con estos, como identificar problemas de calidad, descubrir las primeras ideas sobre los datos y/o detectar subconjuntos interesantes para formar hipótesis con respecto a la información oculta.

1. Recopilar la base datos.
2. Describir los datos.
3. Explorar los datos.
4. Verificar la calidad de los datos.

Los resultados de la fase de comprensión de datos generalmente se documentan en varios reportes que deben redactarse mientras se realizan las tareas respectivas. Los reportes describen además los conjuntos de datos que se exploran. Para el informe final, un resumen de las partes más relevantes es suficiente.

3. Data Preparation

3. Data Preparation

En la preparación de datos se busca construir el conjunto de datos final, el cual se entregará como *input* a las herramientas de modelamiento. A continuación se describen las tareas más comunes a la hora de preparar los datos:

1. Seleccionar los datos.
2. Realizar una limpieza de los datos.
3. Generar variables y construir nuevos datos.
4. Integrar los conjuntos de datos que sean necesarios.
5. Dar formato a los datos finales.

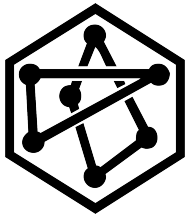
Los reportes en la fase de preparación de datos se centran en los pasos de preprocesamiento que producen los datos que se van a extraer, generalmente se reducen simplemente a comentarios al borde de los códigos para dejar documentado el trabajo.

4. Modeling

4. Modeling

Aquí es donde se seleccionan y aplican varias técnicas de modelamiento, y sus parámetros se calibran a valores óptimos. Por lo general, existen varias técnicas para el mismo tipo de problema de minería de datos y algunas técnicas tienen requisitos específicos sobre la forma de los datos, por lo tanto, volver a la fase de preparación de datos a menudo es necesario.

1. Seleccionar modelos a utilizar.
2. Generar test-train.
3. Construir el modelo.
4. Evaluar el modelo.



Los resultados obtenidos en esta fase se documentan como un reporte corto de los modelos utilizados, anexando el código que se utilizó.

5. Evaluation

5. Evaluation

Es importante evaluarlo a fondo y revisar los pasos ejecutados anteriormente antes de finalizar el proyecto;. un objetivo clave es determinar si hay algún problema importante que no se haya considerado suficientemente. Al final de esta fase, se debe llegar a una decisión sobre el uso de los resultados de la minería de datos.

1. Evaluar los resultados.
2. Revisar los procesos anteriores.
3. Determinar los próximos pasos.

En esta fase se decide si realizar nuevas iteraciones o comenzar la construcción del reporte final.

6. Deployment

6. Deployment

La fase de implementación puede ser tan simple como generar un reporte o tan compleja como implementar un proceso de minería de datos repetible en toda la empresa. En muchos casos, es el cliente, no el analista de datos, quien realiza los pasos de implementación. Sin embargo, incluso si el analista llevará a cabo el esfuerzo de implementación, es importante que el cliente comprenda por adelantado qué acciones deben llevarse a cabo para hacer uso de los modelos creados.

1. Entrega y ejecución del producto.
2. Monitoreo del plan y mantención.
3. Reporte final.
4. Revisar el proyecto.

Este reporte especifica la implementación de los resultados de minería de datos. El reporte final se utiliza para resumir el proyecto y sus reportes anteriores.

II. The Data Science Process

Introduction to Data Science

Patricio Mallea[1]

[1]Data Scientist

May 11, 2020