

巨量資料分析

資料分析實務分享

2017/05/12

Kyper Data Technologies



資料分析步驟

總覽

2

步驟一：總覽全局

步驟二：取得資料

步驟三：發掘與視覺化資料，以取得深入了解

步驟四：修正調整資料，以準備分析

步驟五：分析資料，建立數據模型

步驟六：調整優化模型

步驟七：提出你的見解



總覽全局



總覽全局

了解問題

4

- 了解你的問題領域
- 定義並了解你的問題
 - 機器學習競賽
 - Business Understanding(業務邏輯)
 - 了解企業的痛點跟風險
- 找出衡量的方式



- 預測結果衡量
 - 銷售預測問題: MSE/MAE/R²/MAPE
 - 只能用來比較模型之間的效果
- 如何連結到決策或財務KPI
- 成本效益分析
 - 機器學習競賽對實務的幫助?





提出並檢驗假設



提出並檢驗假設

針對問題以及手上的資料大膽提出假設

8

- 領域知識 + 創意 + 苦工
- 範例: Grupo Bimbo銷售預測 => 考慮影響銷售的變數
 - 天氣
 - 季節性
 - 商品特性
 - 客戶屬性



提出並檢驗假設

以可靠的驗證方式檢驗假設

9

- 方式
 - 視覺化
 - 統計工具
 - 預測模型
- 注意細節
 - 實驗架構
 - 資料品質
 - 可解釋性



提出並檢驗假設

快速迭代

1
0

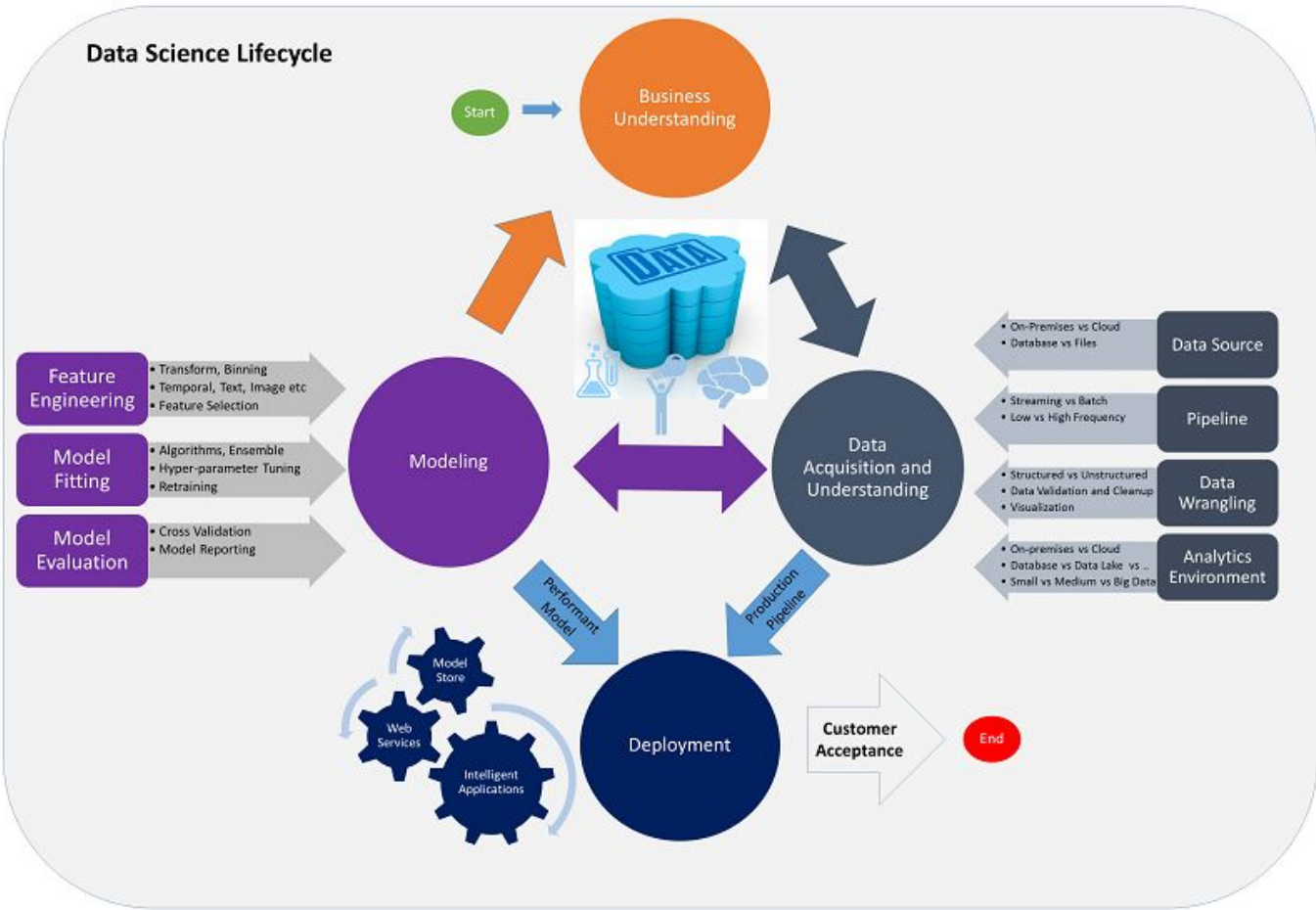
- 建立快速驗證流程
- 提出假設在短時間內驗證
- 盡快發現資料問題或是缺失
- 尋找整合外部資源



專案流程



Data Science Lifecycle



1
2



Reference

1
3

- <https://github.com/Azure/Microsoft-TDSP>
- <https://www.kaggle.com/c/grupo-bimbo-inventory-demand>