Chapter 2 – part 2

Data Analytics Lifecycle

Dr. E. Hamouda
CSCI 398: Introduction to Data Science
Spring 2017

Data Analytics Lifecycle

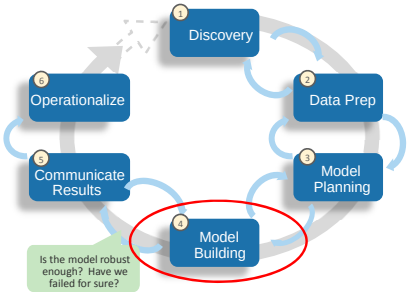
Phase 4: Model Building

- **Develop data sets for testing, training, and production purposes**
  - Need to ensure that the model is valid, sufficiently robust and accurate
    - Accounts for most of the data
    - Has robust predictive power
  - DS develops smaller, test sets for validating approach, training set for initial experiments
- **Get the best environment you can for building models and workflows**
  - fast hardware, parallel processing, etc.

Is the model robust enough?
Have we failed for sure

Model Building

Data Analytics Lifecycle



Phase 5: Communicate Results

Compare outcomes of the modeling to the criteria established for success and failure.

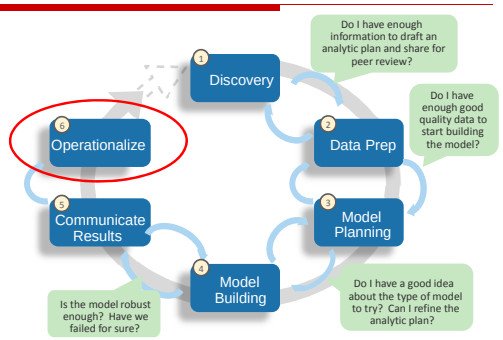
Did we succeed? Did we fail?
The team needs to:

- Summarize the findings, depending on audience
- Interpret the results (include assumptions, limitations of results, etc.)
- Compare to IH's from Phase 1 (prove or disprove?)
- Identify and document key findings
- Quantify business value
- Deliverable: most visible document!!

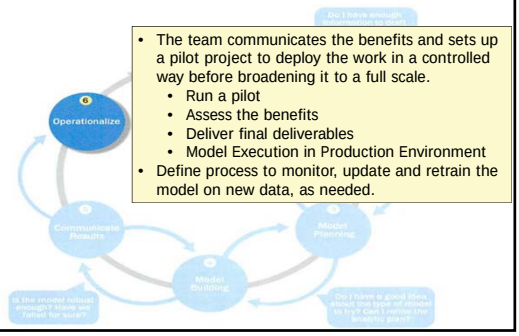
Case Study to Track the Phases in the Data Analytics Lifecycle

Components of Analytic Plan	
Phase 5: Results & findings	• Once customers stop using their accounts for gas and groceries, they will soon erode their accounts and churn. • If customers use their debit cards fewer than 5 times per month, they will leave the bank within 60 days.
Business Impact	• If we target customers who are high risk for churn, we can reduce customer attrition by 25%. • This would save $3 million in lost of customer revenue and avoid $1.5 million

Data Analytics Lifecycle



Phase 6: Operationalize



- The team communicates the benefits and sets up a pilot project to deploy the work in a controlled way before broadening it to a full scale.
  - Run a pilot
  - Assess the benefits
  - Deliver final deliverables
  - Model Execution in Production Environment
- Define process to monitor, update and retrain the model on new data, as needed.

Analytic Plan

Mini Case

Components of Analytic Plan	Retail Banking: Yoyodyne Bank
Phase 1: Discovery Business Problem Framed	How do we identify churn/no churn for a customer?
Initial Hypotheses	Transaction volume and type are key predictors of churn rates.
Data	5 months of customer account history.
Phase 3: Model Planning - Analytic Technique	Logistic regression to identify most influential factors predicting churn.
Phase 5: Result & Key Findings	Once customers stop using their accounts for gas and groceries, they will soon erode their accounts and churn. If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
Business Impact	If we can target customers who are high-risk for churn, we can reduce customer attrition by 25%. This would save $3 million in lost of customer revenue and avoid $1.5 million in new customer acquisition costs each year.

Analyst Wish List for a Successful Analytics Project

- **Data & Workspaces**
  - Access to all the data, including raw data, structured and various states of unstructured data as needed
  - Up-to-date data dictionary to describe the data
  - Area for staging and production data sets
  - Ability to move data between workspace & staging area
  - Analytic sandbox with strong computing power to experiment with the data
- **Tools**
  - Statistical/mathematical/visual software such as SAS, Matlab, R, java tools, Tableau, Spotfire
  - Environment for collaboration with team members
  - Tool/place to log errors with systems or data sets