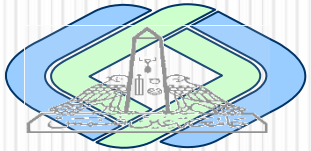
Faculty of Computer
and Information Sciences
Ain Shams University



كلية الحاسبات والمعلومات
جامعة عين شمس

Data Science

Code:

Instructor

Prof. Dr. Abeer M. Mahmoud
Professor of Computer Science-faculty of Computer and Information Sciences-
Ain Shams University
Abeer.mahmoud@cis.asu.edu.eg

Data Science and Big Data Analytics v2

Topics : Data Science and Big Data Analytics Course

Introduction to Big Data Analytics + Data Analytics Lifecycle	Review of Basic Data Analytic Methods Using R	Advanced Analytics – Theory and Methods	Advanced Analytics - Technology and Tools	The Endgame, or Putting it All Together + Final Lab on Big Data Analytics
Big Data Overview State of the Practice in Analytics The Data Scientist Big Data Analytics in Industry Verticals Data Analytics Lifecycle	Using R to Look at Data - Introduction to R Analyzing and Exploring the Data Statistics for Model Building and Evaluation	K-means Clustering Association Rules Linear Regression Logistic Regression Naive Bayesian Classifier Decision Trees Time Series Analysis Text Analysis	Analytics for Unstructured Data (MapReduce and Hadoop) The Hadoop Ecosystem In-database Analytics – SQL Essentials Advanced SQL and MADlib for In-database Analytics	Operationalizing an Analytics Project Creating the Final Deliverables Data Visualization Techniques + Final Lab – Application of the Data Analytics Lifecycle to a Big Data Analytics Challenge

Data analytics lifecycle

Data analytics lifecycle

Upon completing this module, you should be able to:

- ✓ List key phases of the data analytics lifecycle.
- ✓ Apply the data analytics lifecycle to a case study scenario.
- ✓ Describe the purpose of an analytics plan.
- ✓ Name the four core deliverables for a successful project.

Lesson: Data analytics lifecycle overview

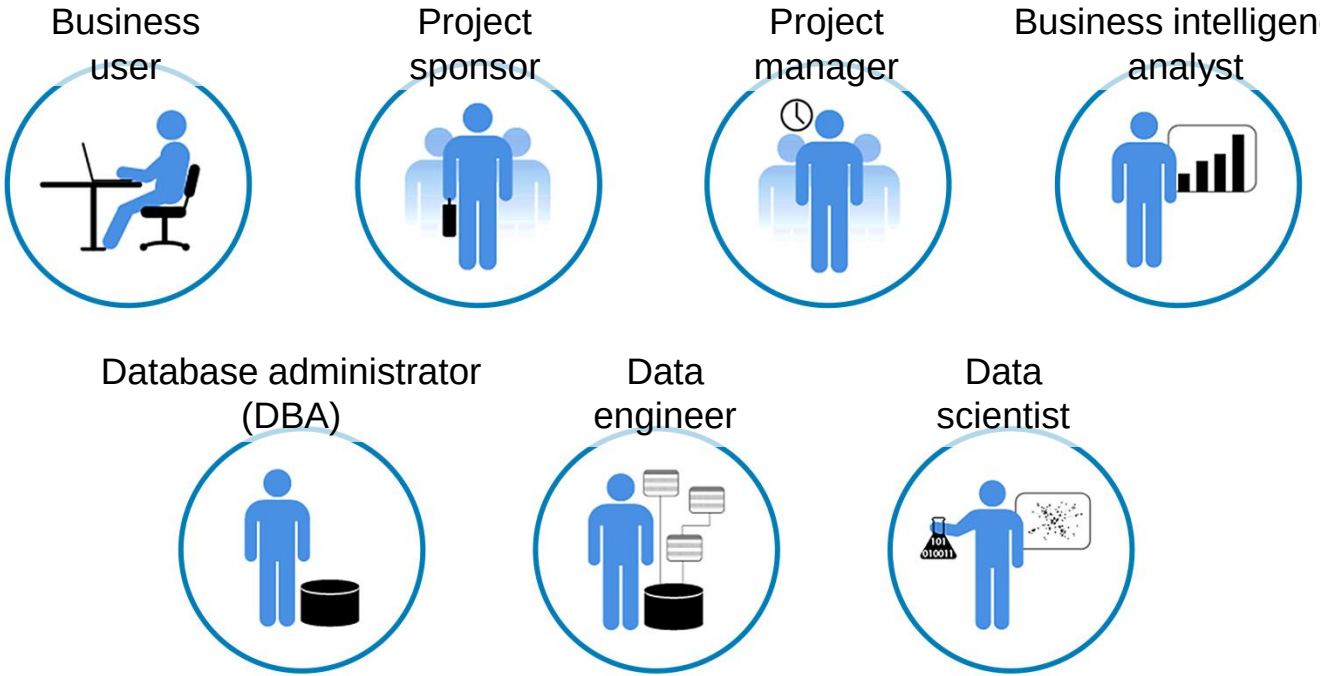
Data analytics problems: What is your approach?

How do you approach your analytics problems?

How do you plan an analytics project to address business problems?

Do you follow a methodology or some kind of framework for an analytics project?

Key roles for a successful analytics project



Key Outputs from a Successful Analytic Project, by Role

Role	Description	What the Role Needs in the Final Deliverables

Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized	• **Sponsor Presentation** addressing:
		  • Are the results good for me?
		  • What are the benefits of the findings?
		  • What are the implications of this for me?

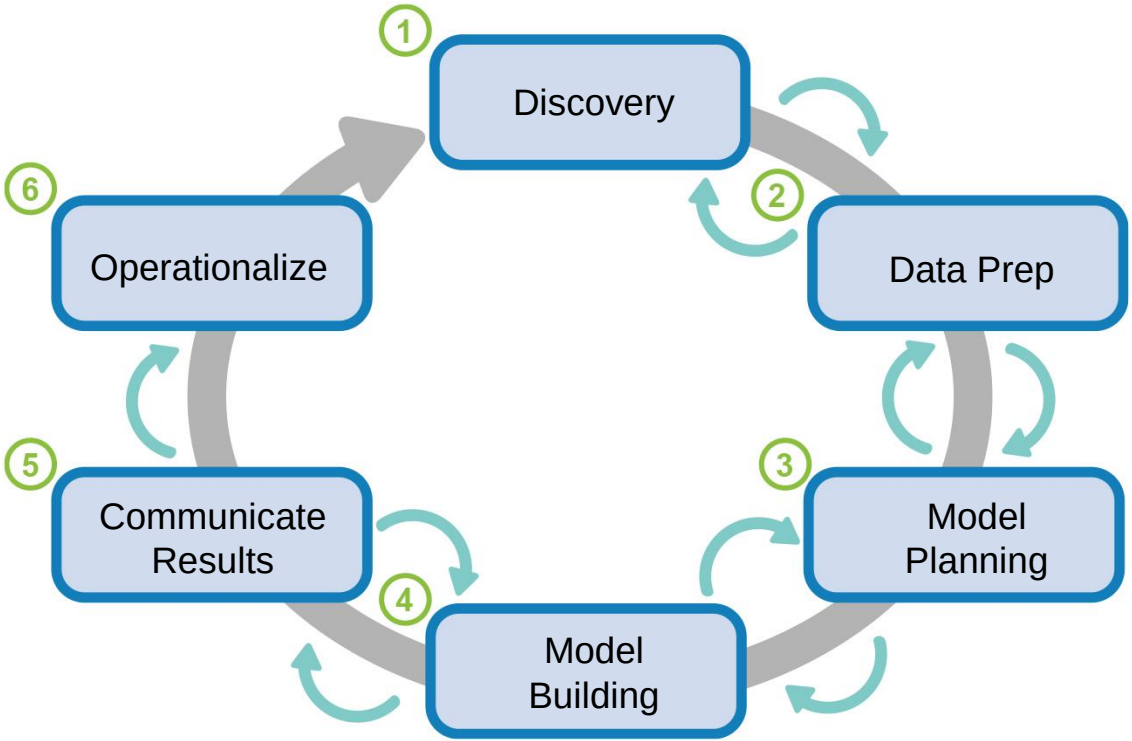
Project Sponsor	Person responsible for the genesis of the project, providing the impetus for the project and core business problem, generally provides the funding and will gauge the degree of value from the final outputs of the working team	• **Sponsor Presentation** addressing:
		  • What's the business impact of doing this?
		  • What are the risks? ROI?
		  • How can this be evangelized within the organization (and beyond)?

Project Manager	Ensure key milestones and objectives are met on time and at expected quality.	
Business Intelligence Analyst	Business domain expertise with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective	• Show the **analyst presentation** • Determine if the reports will change
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management, extraction and support data ingest to analytic sandbox	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of the working team	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Data Scientist	Provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met	• Show the **analyst presentation** • Share the **code**

Why use data analytic lifecycle

- Ensures the business problems are well-defined early in the project:
  - What is the desired business outcome?
  - How will success or failure be determined by the business stakeholders?
- Provides a comprehensive, repeatable method for conducting analyses
- Aids communicating key tasks and assignments within the team
- Plans and scopes the amount of work involved
- Properly sets expectations for the project stakeholders

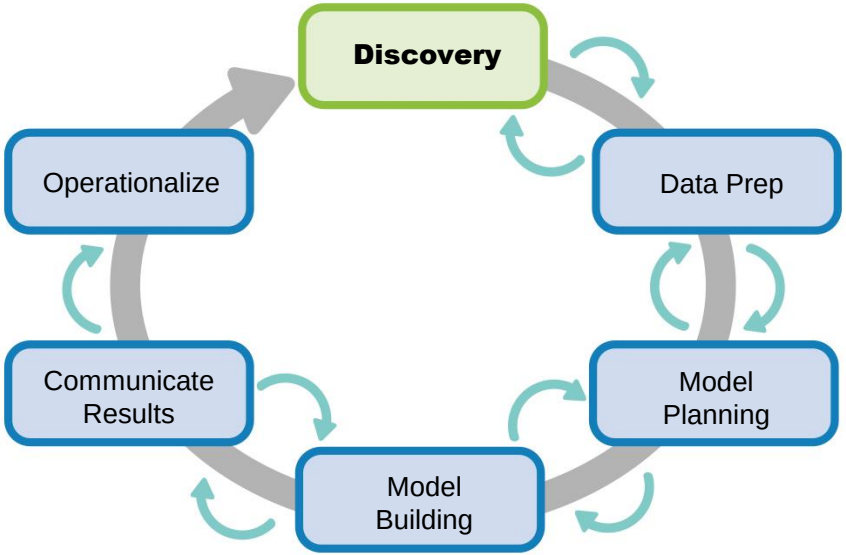
Introduction to data analytic lifecycle



Lesson: Discovery phase

Discovery phase—key activities

- Draft the business problem statement.
- Conduct stakeholder and business expert interviews.
  - Identify the current business state and any pain points.
  - Understand what related projects have been attempted in the past.
  - Refine the business problem statement.
- Reframe the business problem as an analytics challenge.
- Assess resource needs and availability.
- Draft an analytic plan.



Draft the business problem statement

- Clearly articulate the current situation and pain points.
- Determine why addressing the problem matters.
- Include the intended outcome or ideal state.
- Tailor the statement to the key stakeholders.

Discovery—interviewing the project sponsor

Ask about:

- Business problem?
- Pain points?
- Expected outcome?
- If problem not addressed?
- Stakeholder obstacles?
- Constraints?
- Who could provide insights to problem and processes?
- Who has final approvals?

Discovery—interviewing other stakeholders and experts

Ask about:

- Your responsibilities?
- Pain points?
- Related project attempts?
- Existing tools/applications?
- Source systems and available datasets?
  - How to obtain access?
- Status of current IT infrastructure?
- Analytic sandboxes?
- Privacy concerns with datasets?
- Future IT infrastructure upgrades?
- Who else should be interviewed?
- Other insights or recommendations?

Draft an analytics plan

An **analytics plan** is the documentation used to guide the project team through the data analytics lifecycle. It is a living document that will be updated as we progress through different stages in the lifecycle.

Ask about:

- Provides a framework for understanding:
  - What has been accomplished so far
  - How the project should proceed
- Useful to remind the project team about the business objectives and the analytic approach
- As the analysis proceeds, analytic plan updates will often be necessary:
  - It is unlikely that everything will be known at the end of the Discovery phase
  - Scope changes will require additional communications with the project sponsor and stakeholders

Analytics plan template

Component	Phase	Description
Business Problem	1	A concise statement of the current business situation and why addressing the situation matters
Business Impact		The desired result of the project
Analytics Challenge		The business problem expressed as an analytical problem
Data	2	Datasets and variables to be examined
Data Exploration		Assertions about the data to be tested/validated
Proposed Model	3	Analytic techniques to be applied and the model validation approach
Key Findings	5	New insights and the expected business benefits of implementing the project

Test yourself

- Apply the previous template on your graduation project in specific summarized points

Sales analytics project—using data analytics lifecycle

Scenario: The Sales Operations team is struggling to provide the sales managers with useful insights and recommendations on the sales deals currently in the pipeline. The company's executives are frustrated with the constantly changing revenue estimates, particularly at the end of the quarter.

Business Objectives

The Sales Operations team is looking to accurately:

- Identify which sales deals are likely to be completed (booked) in the current quarter.
- Determine which sales deals are at risk.
- Estimate the quarterly revenue.

Key discovery activities—sales analytics project

Interviews with Sponsor and Stakeholders to Understand Pain Points

- Perceived differences of sales deal forecast accuracy by geography and market verticals
- Good-looking deals are suddenly dropped or slip into the next quarter
- Lack of timely information and advanced warning to ensure deals are closed within the quarter

Business Problem Statement

- The lack of timely visibility into which deals are likely to book within the current quarter is preventing the sales team from properly focusing on the correct deals and achieving the quarterly revenue goals.

Key discovery activities—sales analytics project, cont

Analytic challenges

- Design a measure for quantifying the risk in each sales deal
- Enable the sales team with tools to:
  - Better inspect deal risk
  - Help move deals through the sales pipeline
  - Convert high-risk deals to low-risk deals

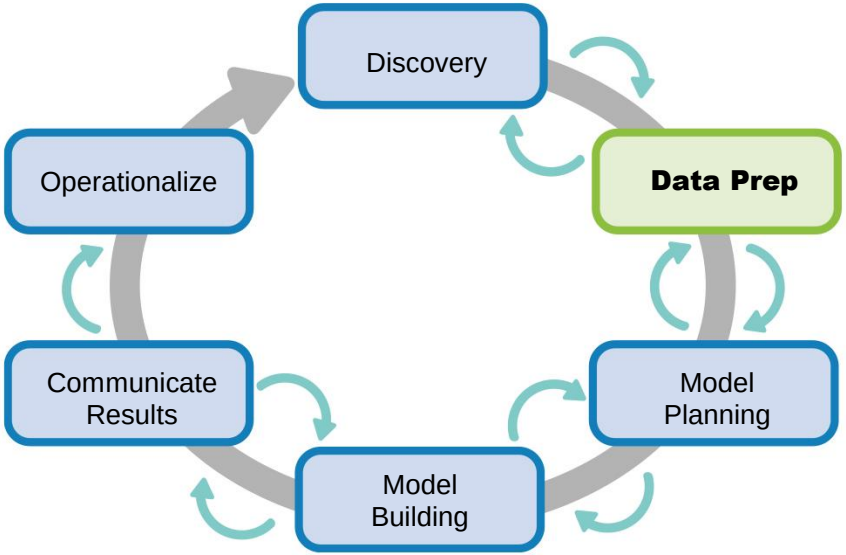
Data sources

- Key data source – Customer Relationship Management (CRM) system to track the sales opportunities
- Other identified data sources provided:
  - Customer demographics
  - Previous product purchases and installations, product service history
  - Service contract renewals
  - Sales representative details (Tenure, Sales Manager, Region)

Lesson: Data preparation phase

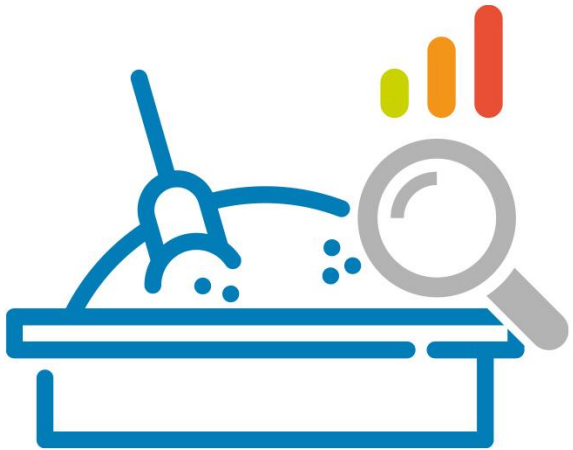
Data preparation—key activities

- Establish the analytic sandbox
- Extract, Transform, Load, and Transform (ETLT)
- Data exploration
- Data conditioning
- Summarize and visualize the data



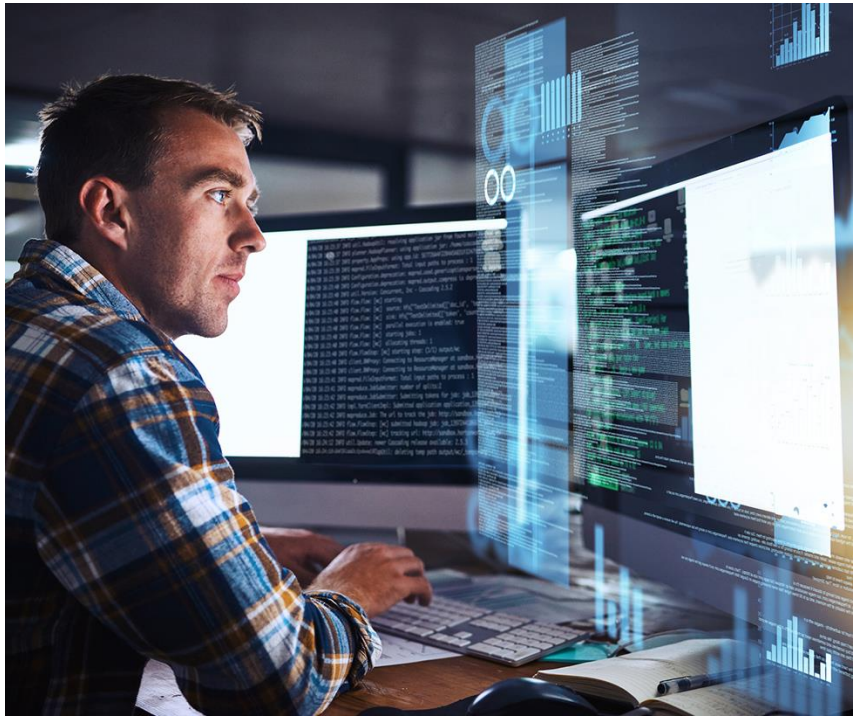
Establish the analytic sandbox

- Estimate the size of the datasets to be loaded.
- Plan for a sandbox about 5 to 10 times the size of the original datasets.
- Consider the data and analytic tools to be utilized.
- Address privacy concerns and security measures.
- Work with IT to prepare the sandbox.



Extract, transform, load, and transform (ETLT)

- Ideally, extract and load the data as stored within the source system.
- Ensure the list of variables included in the sandbox is exhaustive.
- Transform the data within the sandbox.



Data exploration

- Work with business experts to understand what the data elements mean.
- Identify missing data entries.
- Understand how to join datasets or extract the desired information.
- Identify outliers and possible data quality issues.

Data conditioning

- Join and merge the datasets.
- Cleanse the data.
- Normalize datasets.

Summarize and visualize datasets

- Prepare summary statistics and visualizations to better understand the data.
- Identify possible variables that may be related to the outcome to be modeled.

Key data preparation activities—sales analytics project

A Greenplum database was selected as the analytic sandbox:

- Some datasets were already loaded into Greenplum.
- Most of the data was already in another SQL database or well structured.

Data conditioning challenges that needed to be addressed:

- No historical data on the sales deals were maintained.
- Some deals were in parent/child relationships:
  - For example, a deal would ship products to multiple countries or U.S. states.
- Completeness of CRM data was somewhat sales-rep dependent:
  - For example, supplement missing customer data with customer data from other sources

Key data preparation activities—attributes

Data sources:

- Sales opportunities
- Install base
- Service history
- Purchase history
- Sales representative details
- Customer demographics
- Service contracts

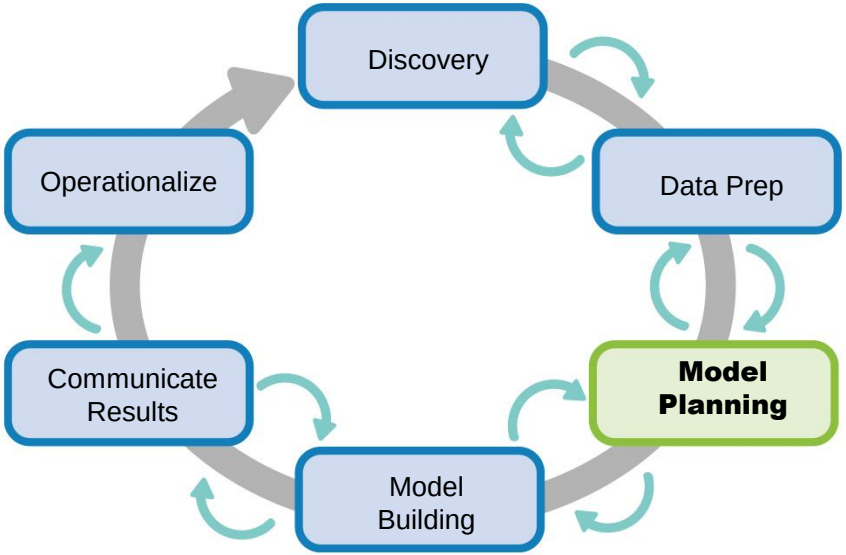
Test yourself

- In the data analytics life cycle , what is meant by Establish the analytic sandbox?

Lesson: Model planning phase

Model planning—key activities

- Variable selection
- Model selection



Variable selection

- Explore the data and understand the relationships among the variables.
- Question and evaluate opinions from stakeholders and domain experts.
- Identify relationships or correlations:
  - Among possible input variables
  - Between input and outcome variables
- Leverage a technique for dimensionality reduction, if applicable.
- Perform additional data transformations to prepare variables for modeling.

Model selection

- Choose an analytical technique or a shortlist of candidates based on:
  - The purpose of the analysis (for example, exploratory or prediction)
  - The types of input and outcome variables (for example, categorical or continuous)
- Decide to fit one model or a series of models:
  - For example, one regression model to handle 50 U.S. states, or 50 regression models—one model for each US state
- Determine the analytic tool to fit the selected model.

Key model planning activities—sales analytics project

The team decides to classify each sales opportunity into one of three categories:

- Book—the sales deal will be booked in the current quarter.
- Push—the sales deal will be deferred into the next quarter.
- Close—the existing opportunity will not result in a booking in any quarter.

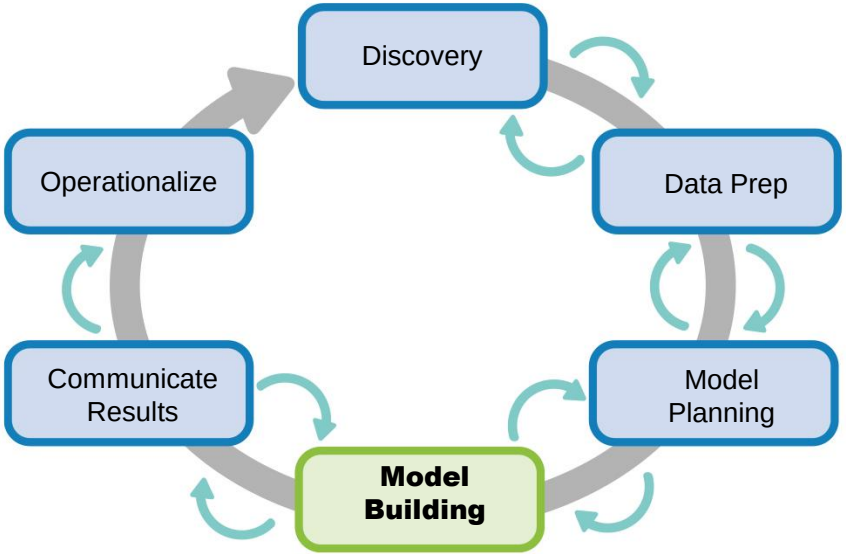
A multinomial logistic regression model was selected:

- The model output provides a probability for each of the three categories.
- The opportunities can then be ranked to determine priorities for the sales team.

Lesson: Model Building Phase

Model building—key activities

- Build training and test datasets.
- Train the selected model:
  - Evaluate the fitted model.
  - Adjust the model accordingly



Build training and test datasets

A typical approach is as follows:

- Randomly select a subset (say 70–90%) of the available data to train the model (**training dataset**).
- Use the remaining data as the **testing dataset** to evaluate the performance of the fitted model.

Additional considerations:

- Apply stratified random sampling to ensure proper representation from all groups.
- Does the modeling technique implicitly account for the training and testing datasets?

Train selected model

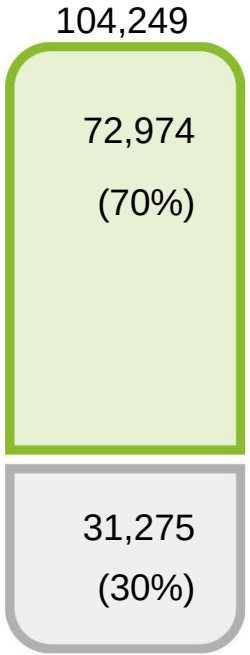
- Identify the useful input variables (feature selection).
- Avoid overfitting the data.
- Verify model assumptions.

Evaluate model

- Is the model accurate enough to meet the goal?
- Are there edge cases that are not properly handled?
- Does the model output/behavior make sense to the domain experts?

Key model building activities—sales analytics project, 1 of 3

One year's worth of sales deals



Training dataset

Randomly selected from base

Testing dataset

Remaining records used to validate model accuracy

Key model building activities—sales analytics project, 2 of 3

Initial model built and validated:

- The multinomial logistic model was trained and key features (input variables) selected.
- The model results were shared with the domain experts.

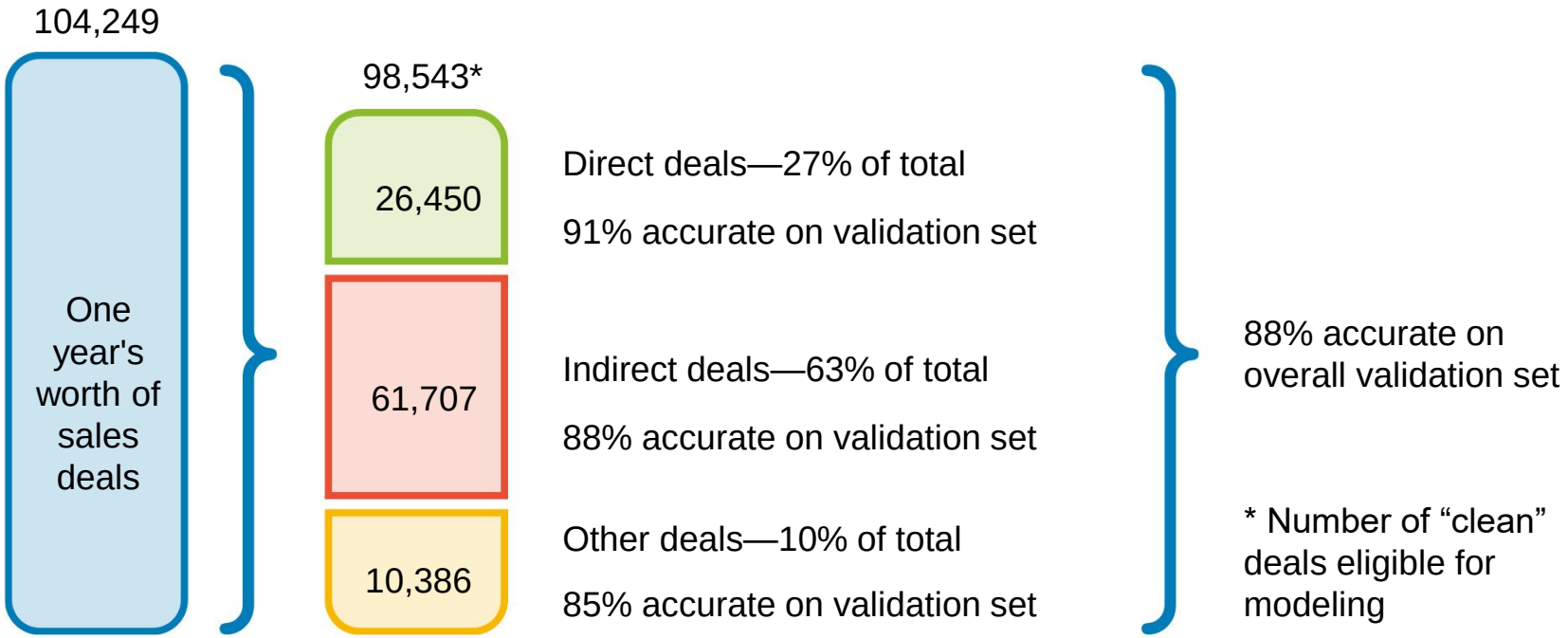
Based on domain experts' feedback:

- The direct sales team will create price quotes well in advance of a deal booking.
- Deals from other teams will not add a quote to the system unless the deal is ready to be booked.
- Thus, attaching quotes is not a reliable predictor for deals through the channel partners.

Sales data stratified into three groups – direct sales, indirect sales, other sales:

- Distinct training and testing datasets were built for each group.
- Three separate multinomial logistic models were built for each group.

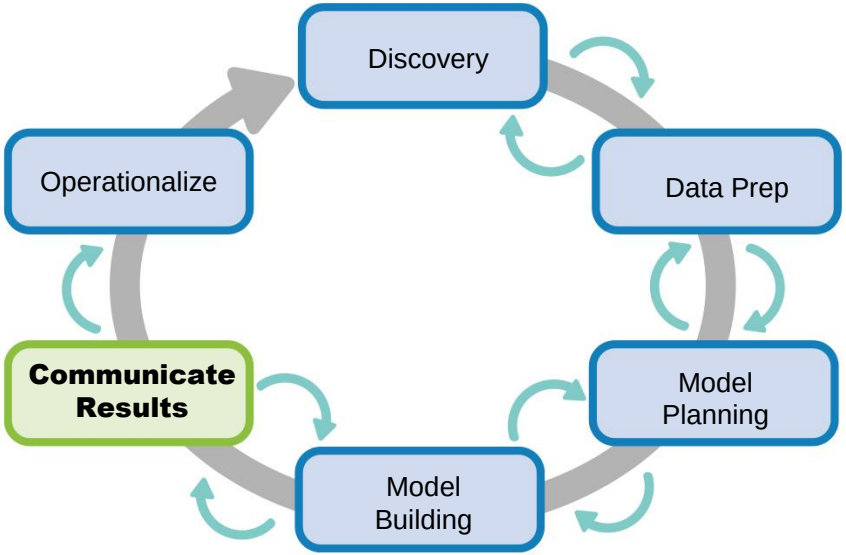
Key model building activities—sales analytics project, 3 of 3



Lesson: Communicate results phase

Communicate results—key activities

- Prepare presentations for sponsors and analysts.
- Share the project results with the various audiences



Share project results with various audiences

- Build a strategy to communicate the findings
- Present the findings to the project sponsor and stakeholders

Core deliverables to meet stakeholders needs

- Presentation for project sponsors and other executives:
  - “Big picture“ takeaways for executive level stakeholders.
  - Determine key messages to aid their decision-making process.
  - Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
  - **Objective:** Demonstrate the business benefit of implementing the model.
- Presentation for analysts:
  - Business process changes
  - Reporting changes
  - Fellow data scientists will want details and technical graphs (for example, ROC curves)
  - **Objective:** Demonstrate the validity of the model.
- Code
- Technical specifications for implementing the code

Key communicate results activities—sales analytics project

i Broad range of opportunity data is collected

ii Data is run through an analytical model

iii Model identifies key input drivers

iv Drivers are used to predict expected deal conversion

Sales opportunity

Product purchases

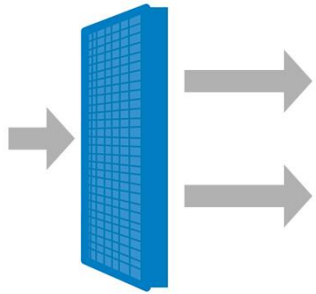
Quote information

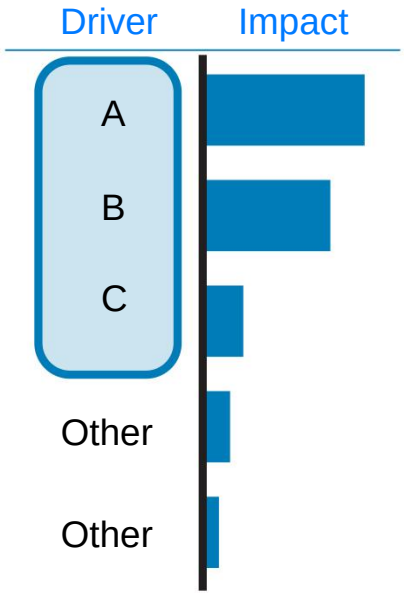
Bookings data

Rep characteristics

Service history

Other relevant data





% Book — Low Risk

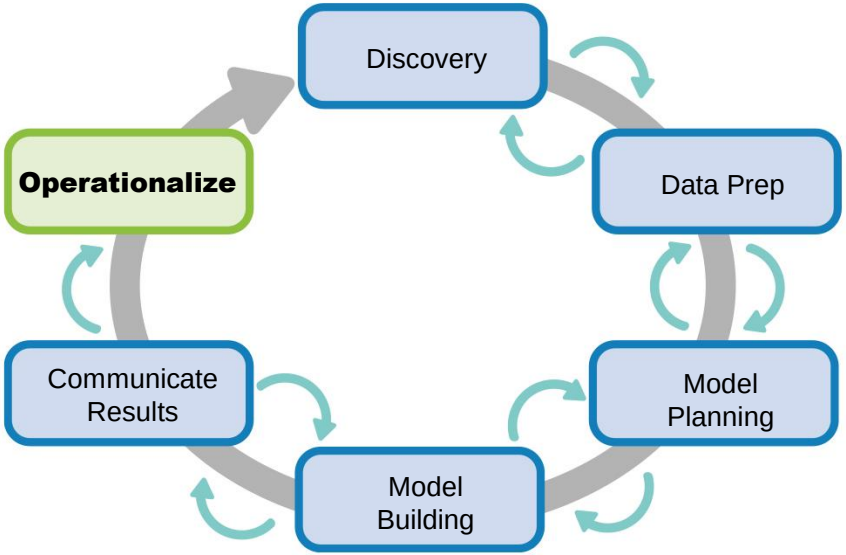
% Push — Med Risk

% Close — High Risk

Lesson: Operationalize phase

Operationalize—key activities

- Provide the code and technical documentation.
- Onboard new team members.
- Deploy the model and monitor its performance.



Provide code and technical documentation

- Complete documentation.
- Obtain acceptance from production owners.

Onboard new team members

- Review history of the project.
- Walk through the analyst's presentation.
- Demonstrate the business value.

Deploy model and monitor

- Deploy the model in test environment.
- Run a pilot project in production.
- Determine some monitoring of the implemented model.
- Recalibrate the model based on feedback.

Key operations activities—sales analytics project

Using the developed models:

- A process was put in place to update the deal predictions weekly.
- The results were packaged into an easily consumable set of documentation.
- The appropriate content was shared with the sales reps, sales managers, regional directors, and so on.

Prepare presentation for sponsors and analysts

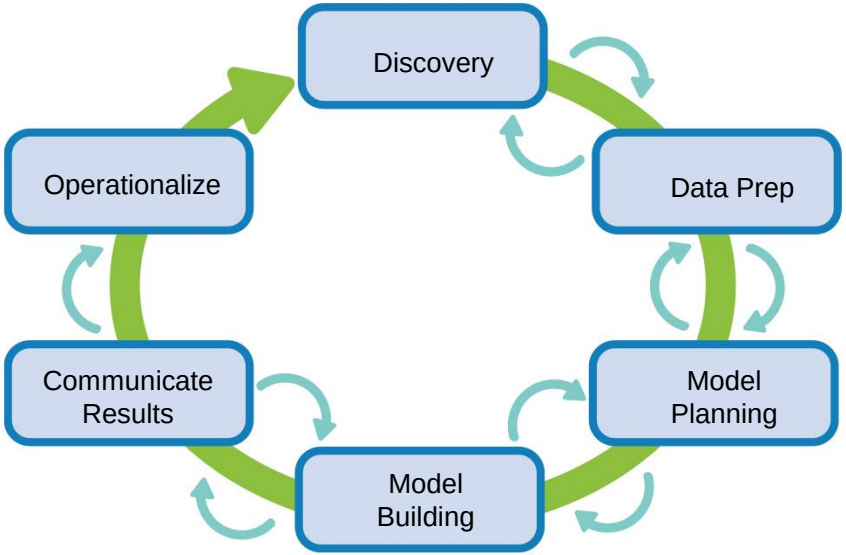
- State the business problem and the objective of the analytic project.
- Provide the business value of implementing the model.
- Prepare recommended next steps.

Lesson: Conclusion

Lifecycle continuation

The work does not end with the operationalize phase:

- The model needs to be monitored and possibly further refined.
- New attributes can be considered for inclusion in the models.
- The delivery mechanism can be simplified with self-service reporting.



Concepts in practice—Pivotal Greenplum Database

- Greenplum Database is an advanced data warehouse.
- Shared-nothing, massively parallel processing design emphasizes parallelism, efficiency, and linear scalability.
- Provides rapid analytics on petabyte-scale data volumes.
- Greenplum Database is based on PostgreSQL.
- Greenplum Database includes features designed to optimize PostgreSQL for analytics workloads.

Check your knowledge

What is a key benefit of implementing ELT (extract, load, and transform) process in the data preparation phase of the data analytics lifecycle?

Check your knowledge

What is a key activity performed in the Communicate Results phase of the data analytics lifecycle?

A. Choosing an analytical technique based on the project goals

C. Sharing the analytic results with stakeholders

B. Data exploration and variable selection

D. Preparing the data repository

Check your knowledge

In which phase does the project team apply clustering, classification, or other analytic techniques?

A. Discovery

C. Model planning

B. Model building

D. Communicate results

Module summary

Key points covered in this module:

- Key phases of data analytic lifecycle
- The importance of a well-defined business problem
- Reframing a business problem as an analytics challenge
- Key deliverables of an analytics project
- The purpose of an analytics plan

Test yourself

- Summarize the case study of sales operation problem in the template of documentation listed in the discovery phase of the data analytics lifecycle